Système de vision hybride : modélisation et application au suivi haute résolution

Julie Badri

N° d'ordre : D.U. 1866
EDSPIC : 410

Université Blaise Pascal - Clermont-Ferrand II

École Doctorale
Sciences Pour l'Ingénieur de Clermont-Ferrand

Thèse présentée par :
Julie Badri

Formation Doctorale STIM :
Sciences et Technologies de l'Information et des Matériaux
Spécialité Automatique et Informatique Appliquée
filière Traitement du Signal et de l'Image

en vue de l'obtention du grade de

Docteur d'Université

spécialité : Vision pour la robotique

Système de vision hybride : Modélisation et application au suivi haute résolution

Soutenue publiquement le 24 octobre 2008 devant le jury :

M. Fabien Feschet	Professeur à l'Université d'Auvergne Président du jury
M. José Luis Lazaro	Professeur à l'Université d'Alcala Rapporteur
M. Frédéric Lerasle	Maître de Conférences au LAAS Rapporteur
M. Christophe Tilmant	Maître de Conférences à l'Université Blaise Pascal Examinateur
M. Quoc-Cuong Pham	Ingénieur de recherche au CEA de Saclay Examinateur
M. Jean-Marc Lavest	Professeur à l'Université d'Auvergne Directeur de thèse

Remerciements

Je remercie tout d'abord Michel Dhome et Thierry Collette pour m'avoir permis de travailler au sein de leur laboratoire respectif, ainsi que mon directeur de thèse Jean-Marc Lavest.

Je remercie Frédéric Lerasle et José Luis Lazaro qui m'ont fait l'honneur d'accepter d'être les rapporteurs de mon travail, ainsi que Fabien Feschet, président du jury, Christophe Tilmant et Quoc-Cuong Pham d'avoir accepté d'examiner ma thèse.

Je remercie particulièrement Christophe et Quoc-Cuong pour leur encadrement et leur soutien au jours le jours durant ces trois années.

Un grand merci à tous les membres du LASMEA et du LSVE pour leur accueil, leur aide et leurs encouragements tout au long de ces trois années. Une pensée particulière à Laetitia du LASMEA, Annie, au *girl power* (Hanna, Laetitia et Hai), Vincent, Pierre, Stevens, Steve et Yoann pour tous ces moments sympathiques autour d'une tasse de thé et leur soutien durant ces longs mois de rédactions.

Enfin, je remercie du fond du coeur mes compagnons de Clermont, les deux François, Mathieu et Julie pour leur soutien, les longues discussions et tous les bons moments passés ensemble ! Merci à vous quatre.

Ma dernière pensée va à mes proches qui ont toujours été là pour moi durant ces trois années.

Résumé

La vidéo surveillance est un sujet scientifiquement complexe dès lors que l'on souhaite intégrer un degré d'automatisation élevé tant dans la détection et la reconnaissance d'un objet spécifique présent dans la scène, que dans la structure de commande rapide et précise du dispositif d'acquisition. Le cadre d'étude de cette thèse se place dans le contexte du suivi automatique de personne à travers la mise en œuvre d'un système de vision hybride (caméra statique et dynamique). Si la littérature est riche dans les dispositifs de vision orientables, les travaux relatifs à la collaboration multi-capteurs sont plus rares dans ce domaine d'application, qui plus est pour des capteurs hybrides alliant un grand champ de vue et une résolution faible avec un dispositif Pan-Tilt-Zoom (PTZ) orientable garantissant le détail de l'information.

Le manuscrit aborde plusieurs problématiques induites par le sujet et donne de premiers résultats de suivi. Une part importante est consacrée à la caractérisation optique et géométrique du dispositif de vision et à la proposition d'une solution originale de calibrage permettant d'estimer une fonction de transfert lorsque les centres optiques des deux capteurs sont pratiquement confondus. L'approche décrite permet d'inférer automatiquement la scène tridimensionnelle observée par le dispositif. Elle établit une relation d'asservissement entre une zone d'intérêt détectée dans l'image grand champ et la commande à appliquer sur le dispositif PTZ pour se focaliser.

Dans une seconde partie, nous abordons la mise en relation des informations extraites de chaque capteur et leur enrichissement mutuel nécessaire à la réalisation du suivi de personne. Après une étude bibliographique, nous détaillons la mise en œuvre d'une approche de filtrage particulaire. L'étape de prédiction du filtre est guidée par la détection issue de la caméra statique et la mesure est donnée par une modélisation d'apparence de la cible extraite de la caméra dynamique. Le formalisme complet du filtre à particules inclut également la loi de commande de la caméra PTZ. Enfin, les premiers résultats de suivi du système complet sont présentés et analysés.

Table des matières

Notations

Les notations suivantes sont écrites de manière générale. En effet, l'indice s ou d sera ajouté à la notation selon que l'on se réfère à la caméra statique ou à la caméra dynamique. Par exemple, le repère caméra associé à la caméra statique est noté $\mathcal{R}_{c_s}$ et celui associé à la caméra dynamique $\mathcal{R}_{c_d}$.

Éléments géométriques

$\mathcal{R}_w$	Repère tri-dimensionnel du monde réel.
$\mathcal{R}_c\left(\vec{\mathbf{x}}, \vec{\mathbf{y}}, \vec{\mathbf{z}}\right)$	Repère tri-dimensionnel lié à la caméra.
$\mathcal{R}_I$	Repère bi-dimensionnel lié au capteur de la caméra.
P_w	Point réel tri-dimensionnel défini dans le repère $\mathcal{R}_w$.
P_c	Point P_w projeté dans le repère $\mathcal{R}_c$.
P_I	Point P_w exxprimé dans le repère $\mathcal{R}_I$.
I	Image provenant du capteur de la caméra.
$\mathscr{C}$	Centre de l'image I.

Éléments mathématiques

$\mathbf{R}$	Matrice de rotation.
$\mathbf{t}$	Vecteur de translation.
$\mathbf{H}$	Matrice d'homographie.
$\mathbf{M}$	Matrice représentant une transformation rigide dans l'espace.

Éléments de la caméra dynamique

α	Angle en azimut (rotation selon le vecteur $\vec{\mathbf{y}}$) de la caméra dynamique (*pan* en anglais).
β	Angle en site (rotation selon le vecteur $\vec{\mathbf{x}}$) de la caméra dynamique (*tilt* en anglais).
Z	Paramètre de grossissement de la caméra dynamique.
Λ	Vecteur de paramètres représentant les paramètres intrinsèques et extrinsèques de la caméra dynamique.

Introduction

Loi n°95-73 du 21 janvier 1995 d'orientation et de programmation relative à la sécurité : articles 10 (Extrait)

« **...** *La transmission et l'enregistrement d'images prises sur la voie publique, par le moyen de la vidéo surveillance, peuvent être mis en œuvre par les autorités publiques compétentes aux fins d'assurer la protection des bâtiments et installations publics et de leurs abords, la sauvegarde des installations utiles à la défense nationale, la régulation du trafic routier, la constatation des infractions aux règles de la circulation ou la prévention des atteintes à la sécurité des personnes et des biens dans des lieux particulièrement exposés à des risques d'agression ou de vol.*

La même faculté est ouverte aux autorités publiques aux fins de prévention d'actes de terrorisme ainsi que, pour la protection des abords immédiats de leurs bâtiments et installations, aux autres personnes morales, dans les lieux susceptibles d'être exposés à des actes de terrorisme.

Il peut être également procédé à ces opérations dans des lieux et établissements ouverts au public aux fins d'y assurer la sécurité des personnes et des biens lorsque ces lieux et établissements sont particulièrement exposés à des risques d'agression ou de vol ou sont susceptibles d'être exposés à des actes de terrorisme.

Les opérations de vidéo surveillance de la voie publique sont réalisées de telle sorte qu'elles ne visualisent pas les images de l'intérieur des immeubles d'habitation ni, de façon spécifique, celles de leurs entrées.

Le public est informé de manière claire et permanente de l'existence du système de vidéo surveillance et de l'autorité ou de la personne responsable. **...** »

Ces dix dernières années, de plus en plus de grandes villes européennes et américaines se sont dotées d'un système de vidéo-surveillance afin de sécuriser au mieux les lieux dits sensibles (voir ci-dessus l'extrait de la loi française). Ce fort développement des systèmes de vidéo-surveillance s'accompagne d'une augmentation de l'activité de la recherche. L'objectif est de développer des systèmes intelligents de vidéo-surveillance qui puissent remplacer la vidéo-surveillance classique (figure 1). En effet, il est humainement difficile pour un opérateur de surveiller simultanément un grand nombre de caméras et ne pas rater un événement qui ne dure que quelques secondes. Ainsi, aujourd'hui, le but des travaux de recherche en vidéo-surveillance est de pouvoir, dans la mesure du possible, accomplir automatiquement des tâches de surveillance.

Si l'on s'intéresse seulement à la surveillance de personnes, celles-ci sont liées aux thématiques suivantes [45] :

1) **Contrôle des accès.** Dans certains lieux de haute sécurité comme les bases militaires, seules les personnes habilitées sont autorisées à entrer. Après la constitution d'une base de données biométriques des personnes habilitées, lorsqu'un visiteur se présente, le système pourra obtenir automatiquement les caractéristiques de la personne telles que

Figure 1: *Centre de surveillance traditionnel et son mur d'écrans.*

sa taille, l'apparence de son visage à partir d'images prises en temps réel et décider si la personne est autorisée ou non à entrer dans le bâtiment.

2) **Identification de personnes.** L'identification des personnes à distance par un système de surveillance intelligent peut aider la police dans la recherche des personnes suspectes. La police peut construire une base de données biométriques des suspects et placer des systèmes de vidéo-surveillance à des endroits où les personnes recherchées ont l'habitude d'être comme, par exemple, les stations de métros, les casinos, etc. Le système doit pouvoir traiter automatiquement les personnes aperçues et juger si elles sont suspectes ou non.

3) **Analyse et statistique des flux et congestion.** En se basant sur la détection de personne, les systèmes de vidéo-surveillance peuvent automatiquement déterminer et analyser le flux de personnes dans des lieux publics tels que les centres commerciaux ou des sites touristiques afin de prévenir les problèmes de congestion.

4) **Détection et alerte en cas d'anomalie.** Dans certaines situations, il est important de pouvoir analyser et déterminer si le comportement d'une personne ou d'un groupe de personne est normal ou non (vol dans un supermarché, agression dans un parking, dégradation de biens ...), si des objets sont abandonnés par des individus, etc. Lorsqu'un comportement suspect est détecté, le système peut alerter la police qui pourra intervenir le plus rapidement possible.

Les premiers systèmes de vidéo-surveillance étaient constitués uniquement d'une caméra statique à focale fixe. En effet, le large champ de vue (figure 2) de ce type de caméra répond bien aux besoins de la vidéo-surveillance. On distingue les caméras perspectives classiques (première colonne de la figure 2) qui présentent, pour les images présentées, un angle de prise de vue d'environ 90° et les caméras avec un objectif fisheye qui ont un angle de prise de vue de l'ordre de 180° (seconde colonne de la figure 2). Par contre, l'inconvénient majeur

de ces capteurs est le manque de résolution, notamment pour les thématiques actuelles de recherche citées précédemment.

FIGURE 2: *Vues issues de différentes caméras statiques dans le cas de scènes en intérieur (première ligne) et en extérieur (seconde ligne) : colonne de gauche, une caméra perspective classique, colonne de droite une caméra statique avec un objectif fisheye.*

Une solution à la faible résolution des caméras à focale fixe est l'utilisation de caméras à focale variable qui permettent l'obtention d'images haute résolution de la scène (première ligne de la figure 3). Comme le champ de vue de ce type de caméra est restreint, notamment dans le cas d'un fort grossissement, les caméras utilisées dans le domaine de la vidéo-surveillance sont des caméras qui peuvent se commander en position (caméra Pan-Tilt-Zoom ou caméra dôme). Ce type de caméra est idéal pour obtenir ponctuellement une image haute résolution d'une zone d'intérêt (figure 3). Par contre, le champ de vue restreint va être un handicap pour effectuer un traitement de suivi. En effet, si la cible a un mouvement brusque, elle risque de sortir du champ de vision de la caméra et aucun algorithme de suivi n'est capable de retrouver la cible perdue, hors champ.

Afin de s'affranchir des inconvénients de chaque type de caméra, de plus en plus de systèmes de vidéo-surveillance se tournent vers la combinaison des deux types de capteurs ([43], [50], [91]) :

- ◇ une caméra **statique** qui permet d'avoir à tout instant une vision globale de la scène,
- ◇ une caméra **dynamique** qui peut à tout instant être pilotée afin d'obtenir une image

FIGURE 3: *Exemples de vues hautes résolutions obtenues avec une caméra dynamique pour un fort zoom de différentes zones d'une scène observée par une caméra statique.*

de plus haute résolution de la cible.
L'intérêt de combiner ces deux types de caméras dans un même système de vidéo-surveillance est de pouvoir effectuer un suivi dynamique haute résolution d'une cible grâce à la caméra dynamique et plus robuste grâce à la vision globale de la caméra statique.

Aujourd'hui, beaucoup de systèmes de vidéo-surveillance utilisent un réseau de caméras composé d'au moins deux caméras. Deux raisons à cela : tout d'abord, l'utilisation de plusieurs caméras permet de remonter à l'information 3D par triangulation et facilite la gestion des occultations, ensuite, un tel système permet de surveiller des environnements plus vastes.

Mais avant de s'intéresser à un réseau de caméras complet, nous avons choisi d'étudier d'abord la brique de base d'un réseau hybride c'est-à-dire une paire de caméras composée d'une caméra statique et d'une caméra dynamique. Plusieurs problèmes se posent : entre autre, comment positionner les caméras l'une par rapport à l'autre, quel type d'information peut-on partager entre les deux caméras et de quelle manière, et comment peut-on faire collaborer les deux capteurs afin de proposer un suivi haute résolution d'une cible.

Contexte de la thèse

L'étude proposée dans ce manuscrit s'inscrit dans le cadre d'une thèse co-financée entre le Laboratoire Systèmes de Vision Embarqués (LSVE) du Commissariat à l'Énergie Atomique de Saclay et le LAboratoire des Sciences et Matériaux pour l'Électronique et d'Automatique

(LASMEA) de l'université Blaise Pascal de Clermont-Ferrand. Ces deux laboratoires ont une thématique de recherche commune qui est le développement de systèmes automatiques d'analyse vidéo temps réel intégrant des techniques évoluées de détection, de localisation, de suivi temporel de personnes dans un environnement ainsi que des méthodes de reconnaissance de gestes ou d'identification de personne.

A plus long terme, ces deux équipes envisagent de coupler à un système de vidéo-surveillance permettant de suivre des objets avec une très bonne résolution un module de reconnaissance de geste ou d'identification de personne. Dans ce contexte de recherche, nous avons donc choisi de positionner les deux caméras de sorte que celles-ci aient le même point de vue afin de faciliter l'ajout d'un module de reconnaissance pour de futurs systèmes plus complexes de vidéo-surveillance. La figure 4 illustre ce choix.

FIGURE 4: *Illustration du positionnement choisi pour les deux caméras composant notre système hybride de vision : à gauche, une caméra statique et à droite, une caméra Pan-Tilt-Zoom.*

La majorité des systèmes de vidéo-surveillance composés de plusieurs caméras utilise l'information 3D, obtenue par triangulation, de position de la cible dans la scène afin de proposer un suivi plus robuste. Nous allons présenter dans ce manuscrit de premiers résultats qui montrent que dans une configuration ne permettant pas d'utiliser ce type d'information, il est tout de même possible de proposer un système de suivi d'objets haute résolution grâce à la mise en place d'une collaboration entre capteurs.

Durant la première partie de cette thèse, nous nous sommes tout d'abord intéressés aux diverses méthodes existantes de calibrage (calibrage fort et calibrage faible) pour un système de vision composé d'une caméra statique et d'une caméra dynamique (chapitre 1). Ensuite, une solution de calibrage, automatique, autonome et adaptée à cette configuration de systèmes de vision, a été développée. Ce calibrage permet d'apprendre une relation liant une position dans l'image de la caméra statique à un couple de paramètres angulaires permettant de centrer la caméra dynamique sur cette position (chapitre 2). Enfin, nous nous sommes attachés à valider notre méthode de calibrage notamment en évaluant, au cours de tests, la précision de la commande de la caméra dynamique estimée (chapitre 3).

Lors de la seconde partie de la thèse, nous avons d'abord abordé la problématique liée au suivi multi-capteurs au travers d'un système maître-esclave, puis en proposant un sys-

tème de collaboration plus étroite entre les deux caméras (chapitre 5). L'élaboration de ces solutions de suivi a été précédée d'un tour d'horizon sur les méthodes classiques de détection et de suivi d'objet en vidéo-surveillance (chapitre 4). Enfin, dans un dernier chapitre (chapitre 6), de premiers résultats de suivi permettant de valider notre proposition de schéma collaboratif entre les deux capteurs sont présentés.

Première partie

Calibrage multi-capteurs hétérogènes

Chapitre 1

Calibrage multi-capteurs : État de l'art

Afin de pouvoir faire du suivi basé sur un système multi-capteurs, on a besoin de définir un modèle permettant de passer d'une information extraite de la caméra statique à une information de commande pour la caméra dynamique. L'obtention de ce modèle fait appel au calibrage de la paire de caméras.

La notion classique de calibrage se réfère à la détermination des propriétés physiques qui sont intrinsèques aux caméras. Cela consiste, dans le cas d'une seule caméra, à obtenir un modèle analytique approché du processus de formation d'images. Dans le cas multi-caméras, le calibrage détermine en plus les relations géométriques décrivant les positions de prise de vue entre les caméras.

Deux approches différentes pour le calibrage d'un système de vision multi-capteurs hétérogènes ont été identifiées :

1. Un calibrage basé sur la connaissance complète du système : les paramètres de chaque caméra et les paramètres inter-caméras sont connus : les paramètres inter-caméras incluent la modélisation du mécanisme de la caméra dynamique. Ainsi, la matrice de passage exprimant la transformation des coordonnées 3D d'un point dans le repère monde en coordonnées 2D dans chaque repère image peut être estimée. On parle de **calibrage fort** du système de vision.

2. Un calibrage considérant le système de vision comme une boîte noire : aucun paramètre de la caméra n'est connu. On estime une relation mettant en correspondance les données image (coordonnées des pixels) de la caméra statique et les paramètres angulaires de la caméra dynamique afin de la centrer sur l'objet. On parlera, en opposition à l'approche précédente, de **calibrage faible**.

1.1 Calibrage *fort* d'un système de vision

Le système de vision à calibrer est un capteur stéréoscopique composé de deux caméras observant la même scène (figure 1.1). On note P_{I_s} le projeté de P_w dans I_s et P_{I_d} le projeté de P_w dans I_d.

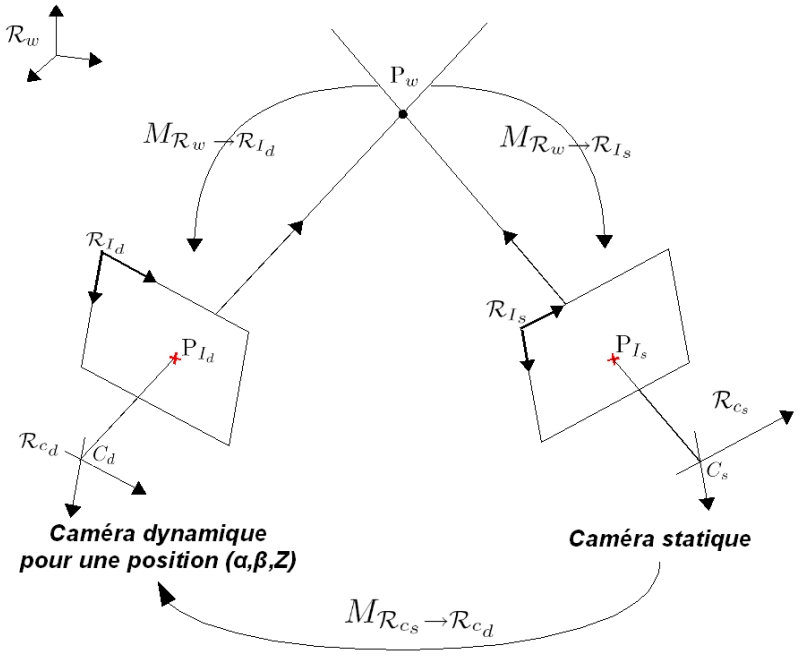


Figure 1.1: *Schéma représentant le système de vision à calibrer ainsi que les relations mathématiques le définissant.*

Le calibrage de ce système de vision consiste à déterminer les matrices de passage $M_{\mathcal{R}_w \to \mathcal{R}_{I_s}}$ et $M_{\mathcal{R}_w \to \mathcal{R}_{I_d}}$. On parle de *calibrage géométrique*. On estime aussi la position et l'orientation relative des deux caméras l'une par rapport à l'autre schématisées par la transformation $M_{\mathcal{R}_{c_s} \to \mathcal{R}_{c_d}}$ sur la figure 1.1. On parle dans ce cas de *calibrage stéréoscopique*.

Le calibrage *fort* est nécessaire si l'on veut effectuer des applications de reconstruction de scène ou de localisation 3D de l'objet cible dans la scène. En effet, lors du calibrage stéréoscopique, on détermine la géométrie des deux capteurs dans le même référentiel. Puis, par triangulation entre les deux capteurs, on remonte à l'information de profondeur. La reconstruction par triangulation est immédiatement liée à l'incertitude sur les primitives et à la largeur de la base stéréoscopique (figure 1.2).

1.1.1 Calibrage géométrique de la caméra statique

Le modèle de caméra utilisé par la suite correspond au modèle de caméra perspective auquel est associé le modèle sténopé (ou *pin-hole*) qui suppose qu'il existe un point C_s,

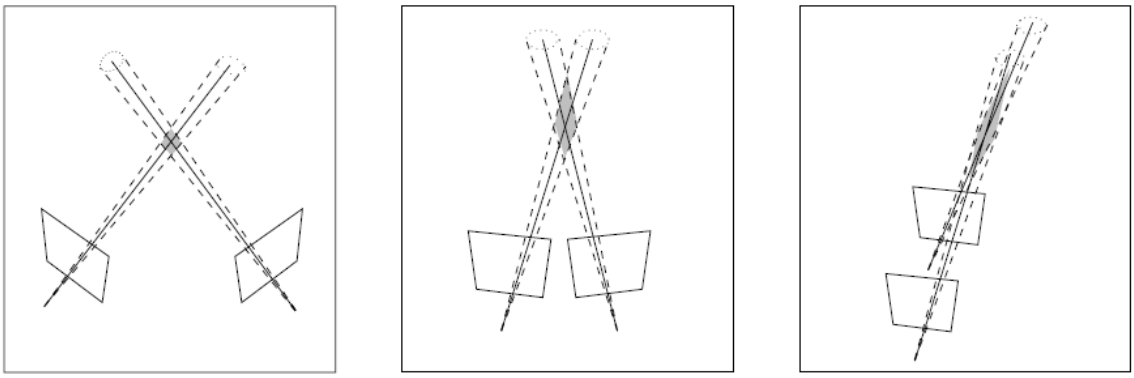


FIGURE 1.2: *Incertitude de la reconstruction [42]. Dans chaque figure, la région ombrée illustre la zone d'incertitude qui dépend de l'angle entre les rayons pour une précision d'appariment donnée. Les points reconstruits sont moins précis lorsque que les rayons deviennent parallèles.*

appelé centre optique de la caméra, par lequel passent tous les rayons lumineux (figure 1.3). Ce centre optique est situé à une distance f (distance focale) du plan image. On appelle axe optique la droite orthogonale au plan image et passant par C_s. Le projeté O de ce point sur le plan image est appelé point principal de l'image.

Le modèle géométrique classique d'une caméra est une projection perspective, résultat de la composition de deux transformations élémentaires représentées par les flèches discontinues sur la figure 1.3.

La première transformation est un changement de repère du point P_w de $\mathcal{R}_w$ en un point P_{c_s} dans $\mathcal{R}_{c_s}$. Elle est composée d'une rotation, $\mathbf{R}$, et d'une translation, t, telles que :

$$\mathrm{P}_{c_s} = M_{\mathcal{R}_w \to \mathcal{R}_{c_s}} \mathrm{P}_w = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \mathrm{P}_w = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0^t & 1 \end{pmatrix} \mathrm{P}_w \tag{1.1}$$

Les paramètres de la transformation rigide entre $\mathcal{R}_w$ et $\mathcal{R}_{c_s}$ ne dépendent que du positionnement de la caméra statique par rapport à son environnement. Ce sont les *paramètres extrinsèques* de la caméra.

La deuxième transformation projette le point P_{c_s} en un point P_{I_s} de I_s et s'écrit [44] :

$$\mathrm{P}_{I_s} = M_{\mathcal{R}_{c_s} \to \mathcal{R}_{I_s}} \mathrm{P}_{c_s} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} k_u & 0 & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} \mathrm{P}_{c_s} \tag{1.2}$$

On suppose une orthogonalité parfaite des lignes et des colonnes de l'image. Les paramètres k_u et k_v sont les facteurs d'échelle vertical et horizontal (en pixel/unité de longueur),

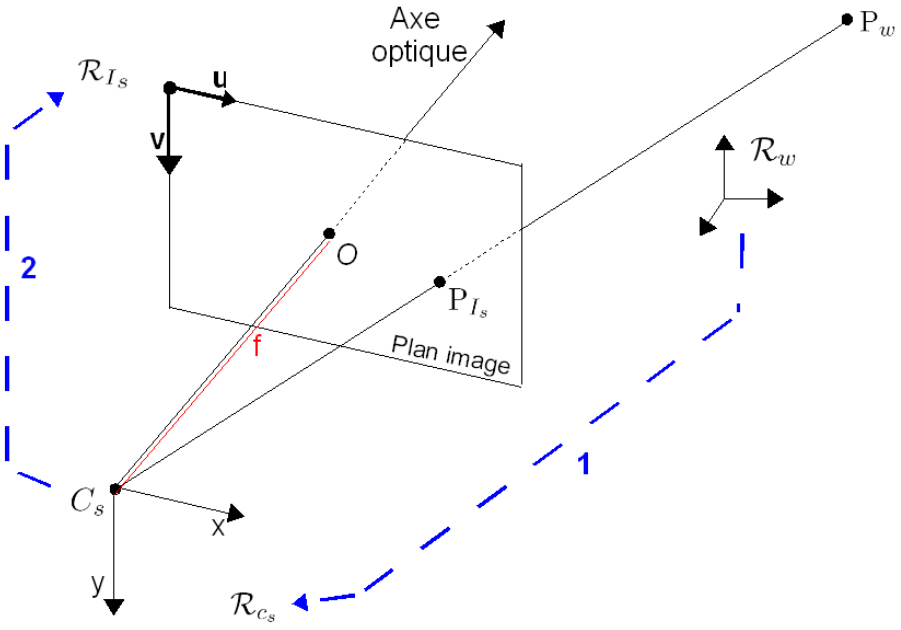


Figure 1.3: *Modèle géométrique d'une caméra : projection perspective, résultat de la composition de deux transformations élémentaires représentées par les flèches bleues discontinues.*

u_0 et v_0 sont les coordonnées du point principal O dans l'image. Ces quatre paramètres sont dépendants de la caméra mais indépendants de la scène 3D. Ce sont les *paramètres intrinsèques* de la caméra.

Le système complet de formation d'images s'exprime alors par la relation suivante :

$$\begin{aligned} \mathrm{P}_{I_s} &= M_{\mathcal{R}_w \to \mathcal{R}_{I_s}} \mathrm{P}_w \\ \mathrm{P}_{I_s} &= \begin{pmatrix} k_u f & 0 & u_0 & 0 \\ 0 & k_v f & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0^t & 1 \end{pmatrix} \mathrm{P}_w \end{aligned} \tag{1.3}$$

En réalité, il faut tenir compte des distorsions géométriques induites par l'objectif non-idéal de la caméra (distorsions dues aux rayons de courbure dans l'image caractérisant les lentilles de l'objectif). En général, on introduit deux termes de distorsions, la distorsion radiale et la distorsion tangentielle modélisées par des polynômes [58].

L'approche classique pour résoudre l'équation 1.3 consiste à estimer les paramètres intrinsèques et extrinsèques de la caméra à partir de la position des points d'une mire connue avec exactitude. La phase de calibrage commence par l'extraction dans l'image des points caractéristiques de la mire de calibrage dont on présente quelques exemples sur la figure 1.4. Plusieurs primitives peuvent être utilisées pour l'extraction de ces points : des croix

avec la méthode de Peuchot [80], des centres d'ellipse avec l'estimateur de Lavest *et al.* [59] ou encore la méthode de Brand [16] et d'autres méthodes décrites dans [12]. Une fois ces points extraits, le processus de calibrage consiste en un problème d'optimisation. La fonction à minimiser est généralement la somme des erreurs quadratiques entre les coordonnées des points extraits dans les images et les coordonnées calculées par le modèle de caméra pour chaque point correspondant de la mire. C'est une fonction non-linéaire par rapport aux paramètres recherchés. La méthode généralement utilisée pour l'optimisation est celle de Levenberg-Marquardt [60].

Figure 1.4: *Exemples de mire de calibrage de caméra.*

1.1.2 Calibrage géométrique de la caméra dynamique

Le calibrage géométrique d'une caméra dynamique est plus complexe que celle d'une caméra statique présentée précédemment. En effet, il faut aussi modéliser le mécanisme de rotation en azimut et en site de la caméra dynamique.

La plupart des méthodes permettant de calibrer une caméra dynamique suppose un modèle cinématique simplifié (figure 1.5) tel que **les axes de rotations soient orthogonaux et alignés avec le centre optique** ([3], [5], [24], [34], [43] [107]). Dans ce contexte, le modèle géométrique de la caméra peut s'écrire de la manière suivante :

$$\mathrm{P}_{I_d} = M_{\mathcal{R}_{c_d} \rightarrow \mathcal{R}_{I_d}} \mathbf{R}_y \mathbf{R}_x M_{\mathcal{R}_w \rightarrow \mathcal{R}_{c_d}} \mathrm{P}_w \tag{1.4}$$

où $\mathbf{R}_x$ représente la matrice de rotation en azimut et $\mathbf{R}_y$ celle en site.

Cette modélisation est suffisante dans le cas d'une caméra dynamique dont le mécanisme de rotation vérifie notamment l'hypothèse que le centre de rotation appartient aux deux axes de rotation, mais elle est insuffisante pour décrire des mécanismes grand public. En effet, il est impossible lors de la production industrielle en masse de caméra pour le grand public d'assurer que le centre optique soit sur les deux axes de rotation à la fois. En pratique, les mécanismes de rotation en azimut et en site sont séparés et par conséquence, il est impossible que le centre optique puisse appartenir aux deux axes de rotation.

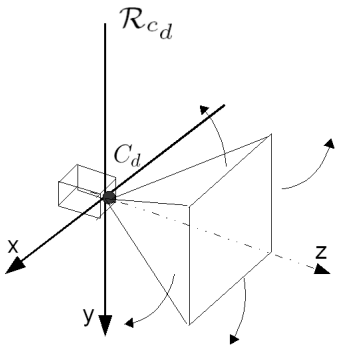


Figure 1.5: *Modèle cinématique simplifié d'une caméra dynamique, [29]. Les axes de rotation en azimut et en site, représentés par des flèches sur la figure, sont modélisés tels qu'ils soient alignés avec le centre optique.*

Afin de mieux modéliser la cinématique d'un mécanisme de rotation quelconque (figure 1.6), il faut donc intégrer ces degrés de liberté dans l'équation 1.4. On obtient la formulation générale suivante, proposée par Davis et Chen [29] puis étendue par Jain *et al.* [50] :

$$\begin{aligned} \mathrm{P}_{I_d} &= M_{\mathcal{R}_{c_d}\rightarrow\mathcal{R}_{I_d}} \mathbf{t}_y^{-1}\mathbf{R}_y\mathbf{t}_y\mathbf{t}_x^{-1}\mathbf{R}_x\mathbf{t}_x M_{\mathcal{R}_w\rightarrow\mathcal{R}_{c_d}}\mathrm{P}_w \\ \mathrm{P}_{I_d} &= M_{\mathcal{R}_w\rightarrow\mathcal{R}_{I_d}}(\Lambda, x, y)\mathrm{P}_w \end{aligned} \tag{1.5}$$

où $\mathbf{R}_x$ et $\mathbf{R}_y$ représentent respectivement la matrice de rotation en azimut et en site, $\mathbf{t}_x$ et $\mathbf{t}_y$ représentent respectivement la translation du centre optique horizontalement et verticalement et Λ représente les paramètres extrinsèques et intrinsèques de la caméra dynamique.

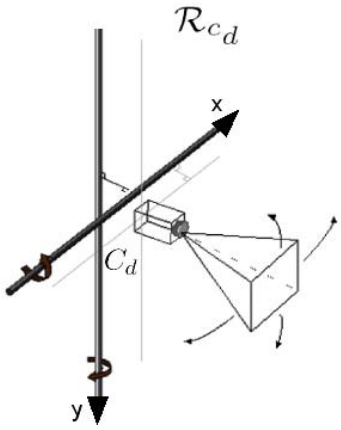


Figure 1.6: *Modèle cinématique généralisé d'une caméra dynamique, [29]. Le mouvement en azimut et en site, représentés par des flèches sur la figure, est modélisé comme des rotations autour d'axes de direction arbitraire dans l'espace.*

Afin de déterminer la transformation $M_{\mathcal{R}_w\rightarrow\mathcal{R}_{I_d}}(\Lambda, x, y)$, Davis et Chen [29] proposent la méthode suivante :

- Un ensemble fini de couples angulaires (α_i, β_i) répartis sur l'ensemble de la plage de mouvement de la caméra dynamique en azimut et en site est défini.
- Pour chaque couple (α_i, β_i), la caméra dynamique est considérée comme une caméra statique et est calibrée selon les méthodes de calibrage géométrique classique.
- Une fois l'obtention d'un jeu de correspondances liant des données 3D à leurs projections 2D dans l'image, les paramètres de la caméra sont estimés par la minimisation de l'équation exprimant la différence entre la projection estimée des données 3D et leurs observations 2D réelles.

A la place d'une mire de calibrage classique, Davis et Chen utilisent une diode électroluminescente afin de couvrir tout le champ de vue de la caméra dynamique du fait de son mouvement.

De façon simultané au calcul de la position et l'orientation des axes de rotation, Jain *et al.* [50] calibrent aussi la correspondance entre les angles requis lors de la commande de la caméra et les angles effectivement réalisés. Les étapes suivantes sont rajoutées à la procédure proposées par Davis et Chen :

- construction par interpolation des expressions $\hat{\alpha} = g(\alpha)$ et $\hat{\beta} = h(\beta)$ reliant les angles α et β requis et les angles $\hat{\alpha}$ et $\hat{\beta}$ effectivement effectués,
- construction par interpolation des transformations t_x et t_y en fonction du zoom : pour un certain nombre de niveaux de zoom prédéfinis, la position entre le centre optique et les axes de rotations sont enregistrés.

1.1.3 Calibrage stéréoscopique

Le calibrage stéréoscopique détermine la matrice de transformation permettant de passer d'un repère caméra à l'autre. D'après la figure 1.1, on peut écrire les équations suivantes :

$$\begin{aligned} \mathrm{P}_{c_s} &= M_{\mathcal{R}_w \to \mathcal{R}_{c_s}} \mathrm{P}_w \\ \mathrm{P}_{c_d} &= M_{\mathcal{R}_w \to \mathcal{R}_{c_d}} \mathrm{P}_w \\ \mathrm{P}_{c_d} &= M_{\mathcal{R}_{c_s} \to \mathcal{R}_{c_d}} \mathrm{P}_{c_s} \end{aligned} \tag{1.6}$$

Les trois transformations sont liées et on peut décrire l'une par rapport aux autres :

$$\begin{aligned} M_{\mathcal{R}_w \to \mathcal{R}_{c_s}} &= M^{-1}_{\mathcal{R}_{c_s} \to \mathcal{R}_{c_d}} M_{\mathcal{R}_w \to \mathcal{R}_{c_d}} \\ M_{\mathcal{R}_w \to \mathcal{R}_{c_d}} &= M_{\mathcal{R}_{c_s} \to \mathcal{R}_{c_d}} M_{\mathcal{R}_w \to \mathcal{R}_{c_s}} \\ M_{\mathcal{R}_{c_s} \to \mathcal{R}_{c_d}} &= M_{\mathcal{R}_w \to \mathcal{R}_{c_s}} M^{-1}_{\mathcal{R}_w \to \mathcal{R}_{c_d}} \end{aligned} \tag{1.7}$$

On remarque que la matrice $M_{\mathcal{R}_{c_s} \to \mathcal{R}_{c_d}}$ se détermine très vite une fois que chaque caméra est calibrée. Cette matrice s'écrit de la manière suivante :

$$M_{\mathcal{R}_{c_s} \to \mathcal{R}_{c_d}} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & b_x \\ r_{21} & r_{22} & r_{23} & b_y \\ r_{31} & r_{32} & r_{33} & b_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1.8}$$

où le vecteur $\mathbf{b} = (b_x b_y b_z)^t$ est le vecteur allant du centre optique de la caméra statique au centre optique de la caméra dynamique (figure 1.1).

1.1.4 Conclusion sur le calibrage fort

Ce type de calibrage donne une connaissance analytique complète de la paire de caméras. Ainsi, connaissant le modèle de projection de chaque caméra et la relation spatiale entre les deux caméras, on peut remonter aux coordonnées 3D d'un point à partir de ses deux projections dans les deux images. Ce type d'information est nécessaire pour la reconstruction de structure 3D, pour la localisation ou la reconstitution de trajectoire. Pour obtenir une bonne précision de la mesure tridimensionnelle, il faut veiller à maximiser l'angle d'orientation des caméras l'une par rapport à l'autre (figure 1.2).

1.2 Calibrage *faible* d'un système de vision

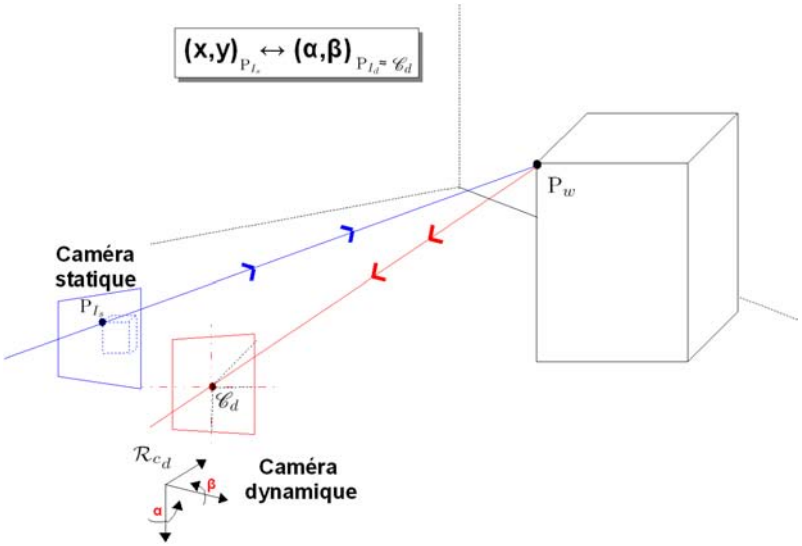


Figure 1.7: *Schéma illustrant le calibrage* faible *d'une paire de caméras.*

Pour des applications de suivi et de reconnaissance, les caméras peuvent être disposées de façon à avoir le même point de vue afin de pouvoir mettre en correspondance les informations issues de chacun des capteurs (figure 1.7). Cette disposition permet d'avoir une approche multi-résolution de la scène. Le problème rencontré, lorsque les capteurs sont relativement proches l'un de l'autre [35], vient de la difficulté à modéliser et estimer les paramètres de mouvement de la tourelle surtout lorsque les angles de mouvement restent faibles. L'approche du calibrage faible permet de contourner cette difficulté. En effet, comme vont l'illustrer les différents travaux cités par la suite, ce type de calibrage permet d'estimer les paramètres de commande de la caméra dynamique à partir d'informations

venant de la caméra statique sans aucune connaissance analytique du système de vision. Cela consiste à déterminer une relation liant les coordonnées des pixels de l'image de la caméra statique et les angles de rotation de la caméra dynamique. Néanmoins, ce type de calibrage est spécifique à la scène observée.

De plus, ce type de calibrage permet de s'affranchir d'une modélisation explicite des distorsions dans l'image dues à l'imperfection du mécanisme optique. En effet, ces imperfections sont implicitement prises en compte dans la relation établie entre la position de P_{I_s} dans I_s et le couple angulaire (α, β) tel que $\mathscr{C}_d$ soit centré sur P_w(figure 1.7).

D'autre part, avec cette approche, le volume 3D de la scène observée est encodé indirectement. En effet, pour une paire stéréoscopique classique, à un point de l'image correspond une droite épipolaire dans l'autre image. Le fait d'utiliser le contenu de l'image basé sur le volume 3D de la scène observée fixe le point sur l'épipole et encode ainsi indirectement le volume 3D de la scène.

Ce type de calibrage a notamment été mis en pratique par Zhou *et al.* [112] et Senior *et al.* [91]. Ces méthodes vont être détaillées.

La méthode proposée par Zhou *et al.* est d'établir une table de correspondance (Look Up Table, LUT) liant les coordonnées des pixels de l'image de la caméra statique avec les angles en azimut et en site correspondants de la caméra dynamique telle que la position de P_w définie dans I_s soit au centre $\mathscr{C}_d$ de I_d. Cette table de correspondance est définie en deux temps :

1. *Établissement de la LUT pour un ensemble fini de points* P_{I_s} *de* I_s :
   Soit $\mathrm{P}^i_{I_s}$ un point de l'ensemble de position prédéfini. Pour chaque point $\mathrm{P}^i_{I_s}$, un point P^i_w de la scène réelle lui correspond. Zhou *et al.* pilotent manuellement la caméra dynamique afin d'aligner $\mathscr{C}_d$ sur P^i_w. Les angles (α_i, β_i) correspondants sont enregistrés dans la LUT et indexés par les coordonnées des points $\mathrm{P}^i_{I_s}$.
2. *Élargissement de la LUT à l'ensemble des pixels de* I_s :
   Pour tout point P_{I_s} de I_s, les angles (α, β) tels que P_{I_d} (correspondant au point P_{I_s}) soit confondu avec $\mathscr{C}_d$ sont obtenus par interpolation linéaire des angles azimut et site des deux plus proches points appris manuellement de P_{I_s}.

L'utilisation d'une méthode d'interpolation linéaire sur des données non-linéaires, dues aux distorsions non corrigées et au volume 3D observé, entraîne une erreur sur l'estimation des paramètres angulaires de commande de la caméra dynamique. Dans leur application, Zhou *et al.* se servent de cette relation pour initialiser le suivi de personne dans la caméra statique, la précision de ce type de calibrage est alors suffisante pour amener la cible dans le champ de vue de la caméra dynamique.

Plus récemment, Senior *et al.* [91] ont proposé une procédure plus automatique que celle de Zhou *et al.* . Pour commander la caméra dynamique, ils utilisent une séquence connue de transformations qui permet de lier une position dans I_s avec les angles de rotation de la caméra dynamique (figure 1.8). Cette méthode est dédiée au suivi de piéton. Ces transformations sont basées sur l'hypothèse que le sol de la scène projeté dans l'image puisse

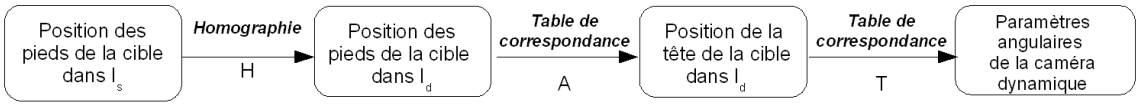


Figure 1.8: *Séquence de transformations permettant à la caméra dynamique de suivre une personne suivie par la caméra statique [91].*

être considéré comme plan. Dans ce cas, une transformation linéaire, en l'occurrence une homographie, est suffisante pour apparier les points du sol de la scène dans I_s avec leurs correspondants projetés dans I_d. Le système apprend automatiquement cette homographie H en utilisant une méthode basée sur celle de Stauffer et Tieu [97]. Les paires de points appariés sont obtenus à partir de vidéo de suivi de cible venant de chaque caméra synchronisé. La détermination de la transformation T, définie sur la figure 1.8, se fait par l'apprentissage automatique d'une table de correspondance entre les angles (α, β) de la commande et le mouvement du centre optique. Pour chaque paire (α, β) d'un ensemble de commandes prédéfinies, Senior *et al.* enregistrent le déplacement du centre optique par rapport à sa position (x_0, y_0) lorsque la caméra dynamique est dans sa position de référence (celle où H a été calculée). Le système décrit une spirale jusqu'à ce que le déplacement du centre optique sorte du champ de vue de la caméra dynamique dans sa position de référence. Ensuite, une minimisation aux moindres carrés est utilisée pour déterminer T tel que :

$$\mathbf{\Theta} = T\mathbf{X} \tag{1.9}$$

où $\mathbf{\Theta}$ représente l'ensemble des couples (α, β) de la base, $\mathbf{X}$ représente l'ensemble des coordonnées $(x_i - x_0, y_i - y_0, 1)$ de points correspondant aux couples angulaires de l'ensemble $\mathbf{\Theta}$.

1.3 Choix du type de calibrage

Le but de notre système n'étant pas de faire de la localisation de personne en terme de mesure, on a choisi de positionner les caméras de manière à ce qu'elles aient un même point de vue proche afin de faciliter la collaboration entre les deux caméras, dans l'optique d'ajouter par la suite une brique d'identification de personne ou de reconnaissance de geste. Ce choix implique qu'un calibrage fort de notre système n'est pas obligatoire. De plus, comme on a pu le voir au cours de l'état de l'art, le calibrage fort est basé sur la connaissance des paramètres intrinsèques des caméras et en particulier de la focale de la caméra. Or, la gamme de caméras utilisée ne nous permet pas de remonter à un contrôle suffisament précis de la focale pour la caméra dynamique. Nous nous sommes donc orientés vers un calibrage faible de notre système de vidéo-surveillance.

Les méthodes de calibrage faible proposées par Zhou *et al.* [112] et Senior *et al.* [91] ont, respectivement, l'inconvénient d'être manuelles et de nécessiter une personne qualifiée pour créer les données constituant la base d'apprentissage. Or, nous souhaitons une solution de calibrage qui soit automatique, autonome et qui s'adapte automatiquement à son

environnement.

Notre méthode de calibrage étant spécifique à la scène observée, elle est par conséquent sensible à de forts changements dans la scène ou dans la position de la caméra statique. Il peut donc être intéressant de pouvoir détecter toute anomalie nécessitant de déclencher un recalibrage de la paire de caméras afin de s'adapter à toute variation du volume 3D de la scène et garder ainsi une bonne précision dans le suivi de personne.

On va donc proposer une nouvelle méthode de calibrage faible répondant à nos objectifs. Notre contribution porte sur les points suivants :

- ◇ **Automaticité et autonomie :**
  Proposition d'une méthode de mise en correspondance automatique basée sur l'information courante de la scène entre un point P_{I_s} dans I_s et les paramètres angulaires et de zoom de la caméra dynamique tel que P_{I_d} soit confondu avec $\mathscr{C}_d$. Cette approche a l'avantage de ne nécessiter aucun apport extérieur (mire, création de données d'apprentissage [91], ...).
- ◇ **Précision :**
  Proposition d'une méthode d'extrapolation des correspondances apprises à l'ensemble des pixels de I_s en intégrant notamment l'ensemble des distortions contenues dans l'image.

Chapitre 2

Calibrage multi-capteurs hétérogènes automatique et autonome

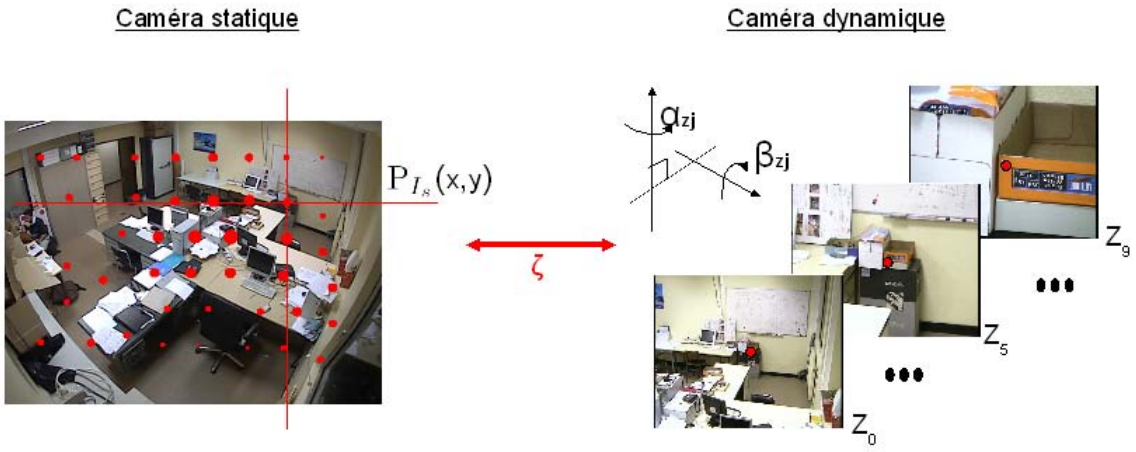


FIGURE 2.1: *Apprentissage de ζ pour un sous-ensemble de positions prédéfinies de I_s (point rouge) pour différentes valeurs du zoom.*

Soit ζ la relation qui pour tout zoom Z établit la correspondance entre les coordonnées (x, y) du point P_{I_s} de I_s et les paramètres angulaires (α, β) tels que P_{I_d} soit confondu avec $\mathscr{C}_d$:

$$(\alpha_Z, \beta_Z) = \zeta(x, y, Z) \; . \tag{2.1}$$

Nous avons choisi de faire correspondre les données en se basant sur des techniques d'asservissement visuel [67] dès lors que la texture locale autour des points d'intérêt est suffisante. Afin de garder un temps raisonnable de calcul, on a choisi de ne pas faire l'apprentissage pour l'ensemble des pixels de I_s. On va donc estimer des correspondances sur un ensemble de pixels ayant un voisinage possédant une texture riche. Dans les zones où il manque de l'information, les correspondances entre les pixels de I_s et les paramètres

angulaires seront déterminées par interpolation. Ainsi, l'apprentissage de ζ se décompose en deux étapes :

1. Mise en correspondance automatique pour un sous-ensemble de positions prédéfinies de I_s par asservissement visuel pour différentes valeurs du zoom (figure 2.1),
2. Mise en correspondance globale par interpolation pour tout pixel de I_s et pour toute la plage de zoom de la caméra dynamique.

2.1 Calibrage multi-capteur par mise en correspondance de données

L'apprentissage de ζ pour un sous-ensemble de positions de I_s est décomposé en deux phases :

1. apprentissage de ζ au zoom initial, noté Z_0, pour toutes les positions prédéfinies,
2. apprentissage de ζ pour différentes valeurs du zoom et pour toutes les positions prédéfinies.

2.1.1 Méthode de calibrage multi-capteur au zoom initial

La méthode de calibrage proposée peut être comparée à l'ICP (*Iterative Closest Point*). La méthode de l'ICP a été introduite par Chen et Medioni [22] et Besl et McKay [7]. Cet algorithme itératif permet de mettre en correspondance deux ensembles de points de manière simple (figure 2.2).

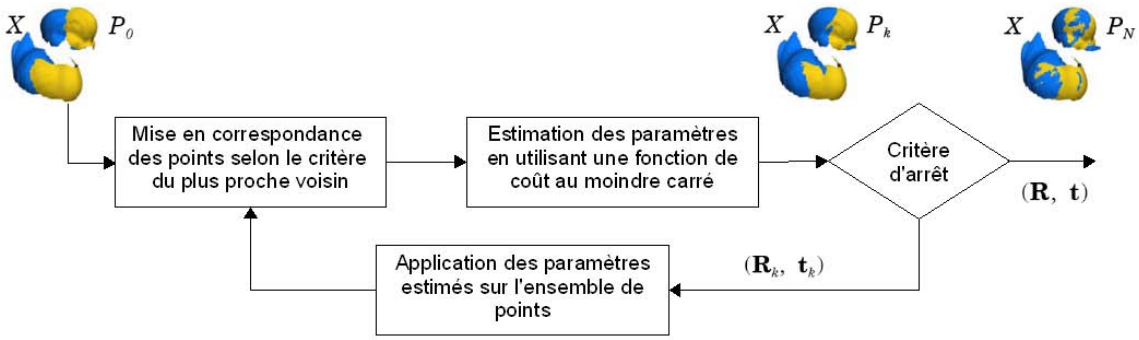


Figure 2.2: *Schéma de l'algorithme initial de l'ICP (Iterative Closest Point) [52].*

Dans notre cas, on cherche à faire correspondre deux ensembles de points issus de fenêtres d'intérêt tels que les paramètres angulaires indiquant la position de la caméra dynamique amènent le point P_{I_d} à être confondu avec $\mathscr{C}_d$.

Les principales étapes de la procédure d'apprentissage de ζ pour un sous-ensemble de points P_{I_s}, illustrées par la figure 2.3, sont les suivantes :

1. Initialisation sur $\mathrm{P}^0_{I_s}$;

2. Pour **chaque point** $\mathrm{P}^i_{I_s}$ de l'ensemble :
   (a) Sélection de la zone d'intérêt issue de I_s, notée I'_s et de I_d
   (b) Extraction et mise en correspondance robuste des points d'intérêt entre I'_s et I_d
   (c) Calcul de la transformation M entre les points d'intérêt de I'_s et I_d
   (d) Calcul des coordonnées de $\mathrm{P}^i_{I_s}$ dans I_d : $\mathrm{P}^i_{I_d} = M \times \mathrm{P}^i_{I_s}$
   (e) Commande empirique de la caméra dynamique afin de diminuer la distance entre les points $\mathrm{P}^i_{I_d}$ et $\mathscr{C}_d$
   (f) Répétition de ce processus sur $\mathrm{P}^i_{I_s}$ tant qu'on ne satisfait pas la condition

$$|\mathrm{P}^i_{I_d} - \mathscr{C}_d| < \epsilon.$$

   Sinon arrêt de la boucle après k itérations
3. Commande de la caméra dynamique pour aller dans le voisinage de $\mathrm{P}^{i+1}_{I_s}$;
4. Retour à l'étape (2) pour traiter le point $\mathrm{P}^{i+1}_{I_s}$;

L'apprentissage de ζ est basé sur la mise en correspondance de données venant de la caméra statique et de la caméra dynamique : on va comparer deux images ayant des résolutions différentes, des points de vue différents et venant de différents capteurs, étape 2b. Ceci nécessite une méthode d'extraction et de mise en correspondance de primitives robuste au changement d'échelle, à la rotation, au changement de point de vue et au changement d'illumination. Mikolajczyk *et al.* [70] comparent différentes méthodes d'extraction et de mise en correspondance de primitives. Ils en concluent que la méthode SIFT de Lowe [63] répond le mieux aux différentes contraintes énoncées précédemment.

Le champ de vue de la caméra dynamique est plus petit que celui de la caméra statique. Afin d'optimiser les résultats de la mise en correspondance des points d'intérêt par la méthode SIFT, une région d'intérêt I'_s est extraite de I_s centrée autour du point $\mathrm{P}^i_{I_s}$ (étape 2b) telle que le champ de vue de I'_s soit équivalent à celui de I_d.

Le but de cet apprentissage est de connaître la correspondance entre les coordonnées (x, y) d'un point $\mathrm{P}^i_{I_s}$ de I_s avec les angles (α, β) de la caméra dynamique tels que l'image I_d soit centrée sur le point $\mathrm{P}^i_{I_d}$. Il est donc nécessaire de pouvoir déterminer la projection des coordonnées (x, y) dans I_d afin d'estimer l'erreur commise. Pour cela, à l'étape 2c, on cherche l'homographie H telle que la mise en correspondance d'un sous-ensemble des points d'intérêt extraits de I'_s et de I_d soit la meilleure. On suppose que localement la distorsion de I_s est négligeable et que l'on peut localement approcher la scène par un plan. Parmi les couples de points mis en correspondance, certains sont erronés. Donc, afin de rendre plus robuste le calcul de H, on utilise la méthode du RANSAC [32] qui fonctionne par tirages aléatoires afin d'éliminer les valeurs aberrantes dans les correspondances de points. Cette méthode d'optimisation est itérative. L'homographie retenue est celle permettant l'appariement du plus grand nombre de points.

Afin de faire converger $\mathrm{P}^i_{I_d}$ vers $\mathscr{C}_d$, on utilise une commande proportionnelle basée sur l'erreur entre les coordonnées de $\mathrm{P}^i_{I_d}$ et les coordonnées de $\mathscr{C}_d$, telle que l'erreur $(\Delta x, \Delta y) =$

$|\mathrm{P}^i_{I_d} - \mathscr{C}_d|$ soit minimisée (étape 2e). Si on considère que les axes de rotation de la caméra dynamique et les axes des coordonnées sont alignés dans le cas de petits déplacements, on peut alors écrire :

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}K_{x\to\alpha} & 0\\ 0 & K_{y\to\beta}\end{pmatrix}\begin{pmatrix}\Delta x\\ \Delta y\end{pmatrix} \tag{2.2}$$

Pendant la phase d'apprentissage, on suppose que la scène 3D est invariante. De plus, n'ayant pas de contrainte temporelle sur la rapidité de la commande et la précision du système étant dépendante du mécanisme de la caméra, une commande proportionnelle suffit dans notre cas.

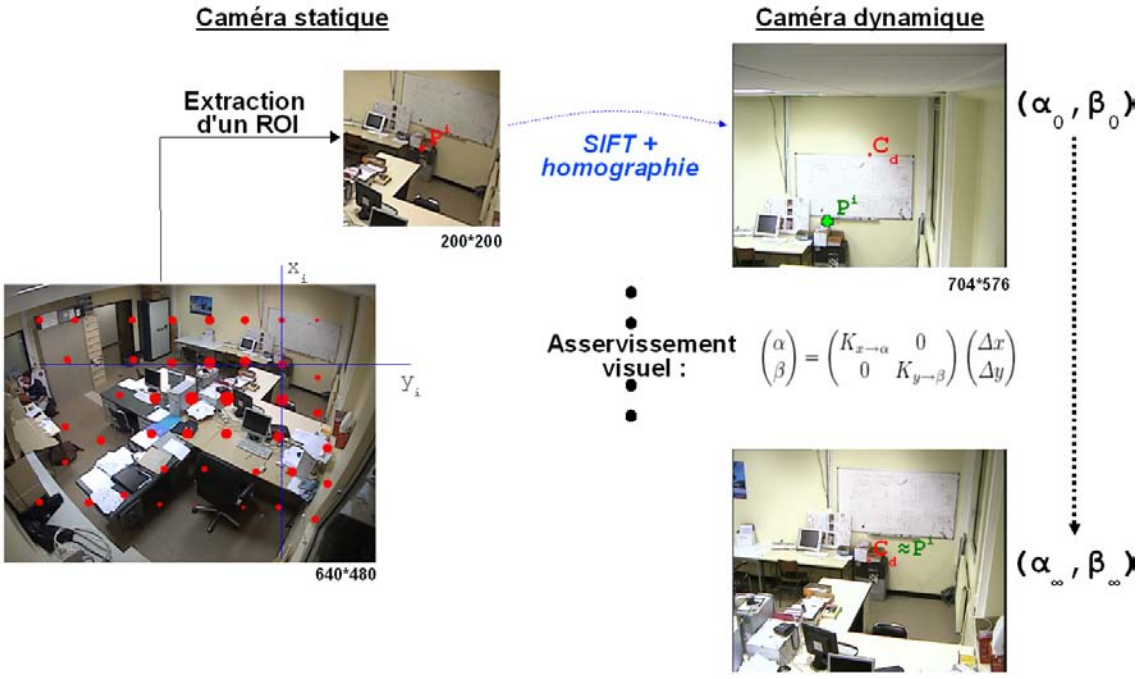


Figure 2.3: *Schéma de l'apprentissage de ζ pour un sous-ensemble de points P_{I_s} à un zoom donné par asservissement visuel.*

2.1.2 Procédure pour amener un point de la scène 3D dans le champ de vue de la caméra dynamique

Cas du point $\mathrm{P}^0_{I_s}$:

Dans le cadre de l'automatisation de la procédure, on a besoin de traiter de façon particulière le premier point de correspondance. L'objectif à atteindre est que le projeté d'un point P_w, correspondant au point $\mathrm{P}^0_{I_s}$, et son voisinage soient visibles dans l'image de la caméra dynamique (étape 1). Pour cela, on tire aléatoirement des paramètres angulaires

jusqu'à ce que I_d ait suffisamment de concordance au sens de la mise en correspondance de points avec le voisinage de $P^0_{I_s}$.

Soit $P^{\mathscr{C}_d}_{I_s}$ le point correspondant à $\mathscr{C}_d$ dans I_s.

Les principales étapes de la procédure sont les suivantes :
Tant que $P^{\mathscr{C}_d}_{I_s}$ n'est pas dans le voisinage de $P^0_{I_s}$

1. Tirage aléatoire d'un couple de paramètres (α, β)
2. Sélection des images courantes I_s et I_d
3. Extraction et mise en correspondance robuste des points d'intérêt entre I_s et I_d
4. Calcul de l'homographie H entre les points d'intérêt de I_s et I_d
5. Calcul des coordonnées de $\mathscr{C}_d$ dans I_s : $P^{\mathscr{C}_d}_{I_s} = H \times \mathscr{C}_d$

Cas des points prédéfinis $P^i_{I_s}$ *:*

Supposons que l'on a déjà traité m points lors de l'apprentissage.

Pour diriger la caméra dynamique dans le voisinage de $P^{m+1}_{I_s}$, on estime le couple de paramètres $(\alpha_{m+1}, \beta_{m+1})$ en se basant sur la connaissance des correspondances apprises précédemment. On recherche parmi les points appris précédemment le point le plus proche de $P^{m+1}_{I_s}$ pour chaque coordonnée de manière indépendante. Ceci donne une bonne initialisation pour un processus itératif afin de raffiner les données.

Si on ne trouve pas de point déjà appris suffisamment proche, on applique la même procédure d'initialisation que pour le point $P^0_{I_s}$.

2.1.3 Méthode de calibrage multi-capteur pour toute la plage de zoom

L'apprentissage de ζ pour différentes valeurs du zoom pour tous les points du sous-ensemble prédéfini est basé sur le même processus que l'apprentissage au zoom Z_0. La différence se situe sur le choix de l'image de référence. En effet, dans le cas précédent, on comparait une image issue de la caméra statique avec une image issue de la caméra dynamique. Lors de l'apprentissage pour différentes valeurs du zoom, on va comparer une image issue de la caméra dynamique à un zoom donné, noté Z_k, avec une image issue de la même caméra à un zoom, noté Z_j, tel que $Z_k < Z_j$. Si l'on avait choisi de se référer à l'image au zoom précédent, l'erreur commise s'additionnerait au fil du processus. En choisissant de se référer à l'image au plus petit zoom permettant de comparer les deux images, on minimise l'erreur que l'on peut commettre au fil du processus.

Les principales étapes de cette méthode illustrées sur la figure 2.4 sont les suivantes :

1. Pour **chaque point** $P^i_{I_s}$:
   (a) Initialisation de la caméra dynamique au zoom de référence Z_{ref} (au départ, $Z_{ref} = Z_0$) en appliquant la commande (α, β) apprise précédemment
   (b) Sélection de l'image courante, noté I^{ref}_d
   (c) Pour **chaque zoom** Z_j tel que $Z_j < Z_{max}$:

i. Application de la commande de zoom Z_j

ii. Extraction et mise en correspondance robuste des points d'intérêt entre I_d^{ref} et I_d^j

iii. Calcul de l'homographie H entre les points d'intérêt des images I_d^{ref} et I_d^j

iv. Calcul des coordonnées du centre $\mathscr{C}_d$ de I_d^{ref} dans I_d^j :

$$\mathrm{P}^i_{I_d^j} = H \times \mathscr{C}_d$$

v. Commande de la caméra dynamique afin de minimiser la distance entre $\mathrm{P}^i_{I_d^j}$ et $\mathscr{C}_d$

vi. Retour à l'étape 1(c)i tant que l'on ne satisfait pas la condition

$$|\mathrm{P}^i_{I_d^j} - \mathscr{C}_d^j| < \epsilon.$$

vii. Si il y a échec du processus, retour à l'étape 1a avec $Z_{ref} \leftarrow Z_{ref+1}$

(d) Retour à l'étape 1c pour traiter le point $\mathrm{P}^i_{I_s}$ au zoom Z_{j+1}

2. Retour à l'étape 1 pour traiter le point $\mathrm{P}^{i+1}_{I_s}$ avec $Z_{ref} = Z_0$

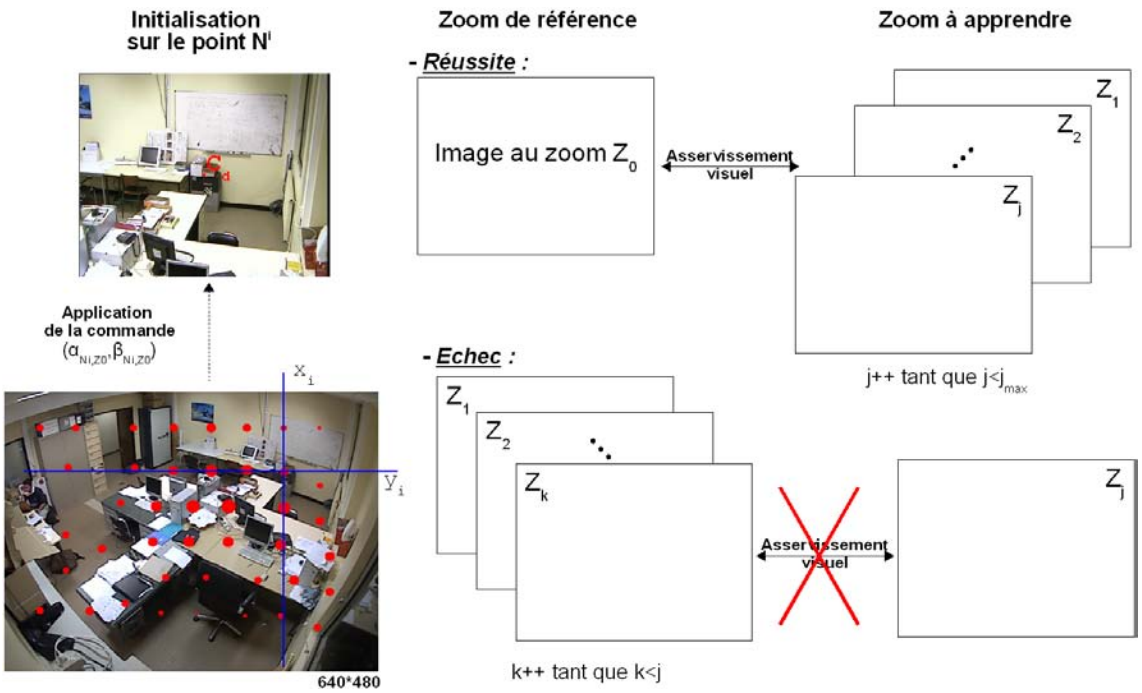


Figure 2.4: *Schéma de l'apprentissage de ζ pour un sous-ensemble de points de I_s pour différentes valeurs de zooms par asservissement visuel.*

2.2 Maillage de l'apprentissage

Lors du processus de calibrage du système de vision, la comparaison des images est basée sur la mise en correspondance de points d'intérêt extraits avec la méthode SIFT de Lowe [63]. Si l'on choisit comme ensemble d'apprentissage un quadrillage régulier sans lien avec l'information 3D de la scène, certains points du quadrillage vont se retrouver dans des zones homogènes. Comme on a choisi d'utiliser l'algorithme SIFT, il n'y aura pas ou peu de points détectés dans ces zones homogènes. L'estimation de l'homographie sera bruitée et le calibrage induira d'importantes erreurs de commande dans ces zones.

C'est pourquoi, on a choisi un quadrillage qui soit adapté à l'utilisation de SIFT dans notre processus. Pour cela, notre quadrillage est composé d'un nombre de points dépendant de la scène 3D afin d'optimiser les résultats d'appariement des points d'intérêt issus de la méthode SIFT. La procédure de création d'un sous-ensemble de points de I_s pour l'apprentissage de ζ est la suivante.

Pour une image I_s donnée, on extrait les points d'intérêt avec la méthode SIFT. Pour chaque pixel de I_s, on affecte une densité de probabilité basée sur les points d'intérêts extraits précédemment basée sur la méthode des fenêtres de Parzen [79] (image centrale de la figure 2.5). La taille de la fenêtre dépend de la relation entre les champs de vue de la caméra statique et de la caméra dynamique au zoom initial. La taille de cette fenêtre sera la même que la taille de la sous-image extraite lors du calibrage du système de vision au zoom Z_0 (partie 2.1.1).

On impose deux contraintes aux pixels qui formeront l'ensemble des points d'apprentissage (figure 2.1) :

1. Rejet des points venant du bord de l'image,
2. Répartition maximale sur toute l'image.

Pour satisfaire la première contrainte, pour une bande B_l de largeur l pixels, la densité de probabilité des pixels de I_s appartenant à B_l, pour chaque bord de I_s, est mise à zéro. Pour notre application, la largeur l vaut 50 pixels.

Ensuite, le premier point du quadrillage correspond au pixel de I_s ayant la plus grande densité de probabilité. Afin de garantir une bonne répartition des points, la densité de probabilité des pixels appartenant au voisinage du premier point est mise à zéro. Dans le cas d'un quadrillage régulier, le voisinage entre deux points correspond au pas séparant deux points du quadrillage.

Cette procédure est répétée jusqu'au dernier pixel de I_s ayant une densité de probabilité non-nulle. Le résultat est illustrée par l'image de droite de la figure 2.5.

2.3 Interpolation des données par *Thin Plate Spline*

La méthode de calibrage proposée permet d'apprendre pour un ensemble de pixels prédéfinis de I_s les paramètres angulaires correspondant de la caméra dynamique tels que cette

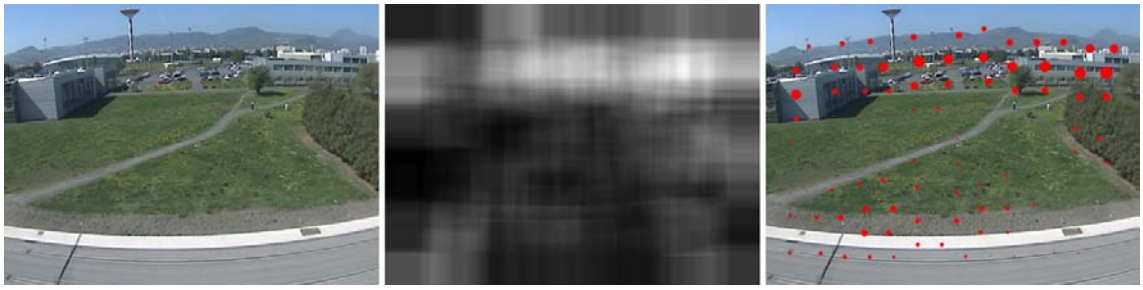


Figure 2.5: *Illustration du résultat de la création d'un quadrillage adapté à la scène 3D : à gauche la scène originale, au milieu le résultat des fenêtres de Parzen et à droite le quadrillage résultant.*

dernière soit centrée dessus. Afin d'étendre cette connaissance à l'ensemble des pixels de I_s, on va interpoler les correspondances déjà connues aux zones de l'image I_s non apprises.

La méthode des *Thin Plate Spline*, d'abord proposé Bookstein [14], est une procédure adaptée pour notre application. L'auteur propose une méthode algébrique pour décrire les déformations spécifiées par deux ensembles de points homologues de l'espace euclidien. Cette méthode produit une fonction d'interpolation f qui transforme le premier ensemble de points, l'ensemble source, vers le second, l'ensemble cible. De plus, cette fonction f est définie partout dans l'espace euclidien et en particulier, dans un voisinage des points source, de sorte qu'elle peut être appliquée à un point quelconque de « l'espace source »pour trouver son homologue dans « l'espace cible ».

Dans notre cas, on cherche la transformation qui permet de passer de l'ensemble des coordonnées des points du quadrillage appris à l'ensemble des angles azimut et site correspondants. Cette transformation va permettre d'estimer ζ (équation 2.1) pour tout pixel de l'image de la caméra statique et pour toute la plage de zoom de la caméra dynamique.

On note r_{ij} la distance euclidienne entre deux points $\mathrm{P}^i_{I_s}$ et $\mathrm{P}^j_{I_s}$ de l'ensemble source. Dans ce cas, la fonction f est la somme d'une partie affine déterminant son comportement à l'infini et d'une partie non linéaire asymptotiquement horizontale :

$$f(x,y) = a_1 + a_x x + a_y y + \sum_j w_j U\left(|\mathrm{P}^j_{I_s} - (x,y)|\right) \tag{2.3}$$

où

- la fonction de base U est la solution fondamentale de l'équation biharmonique $\Delta^2 U = \delta(0,0)$, δ désignant la fonction de Kronecker. La fonction U est dépendante de la dimension de son espace de définition. Ainsi, en 3D, la fonction U est $U(r) = |r|$, tandis que $U(r) = r^2 \ln r$ en 2D et $U(r) = |r|^3$ en 1D,
- les coefficients $\mathbf{a} = (a_1, a_x, a_y)^t$ et $\mathbf{w} = (w_1, w_2, \dots)^t$ sont les solutions du système linéaire :

$$\begin{cases} \mathbf{Kw} + \mathbf{Pa} = \mathbf{v} \\ \mathbf{P^t w} = 0 \end{cases} \tag{2.4}$$

où la matrice $\mathbf{K}$ est une matrice carrée dont la dimension est égale au nombre de correspondances apprises de terme général $U(r_{ij})$, la matrice $\mathbf{P}$ est définie par l'ensemble des points sources et $\mathbf{v}$ est un vecteur contenant un paramètre de l'ensemble cible. Par exemple, $\mathbf{v} = (\alpha_1, \alpha_2, \dots)$ implique que la fonction f soit définie par l'équation 2.3 doit être exprimée pour $f_x(x, y)$ et $f_y(x, y)$ et que le système linéaire 2.4 doit être résolu pour chaque paramètre angulaire α et β.

La popularité de *Thin Plate Spline* provient d'un certain nombre d'avantages. Cette méthode a l'avantage de proposer une transformation d'influence globale et naturellement lisse (figure 2.6) avec des dérivés de n'importe quel ordre. Ce modèle ne possède pas de paramètres qui nécessitent des réglages manuels. Ensuite, la fonction d'énergie liée à cette méthode d'interpolation a une explication physique. De plus, cette méthode d'interpolation permet de pondérer la confiance que l'on a dans les données d'apprentissage de l'interpolation : on peut imposer de garder lors de l'estimation de la transformation une correspondance exacte entre les données utilisées pour l'apprentissage de celle-ci.

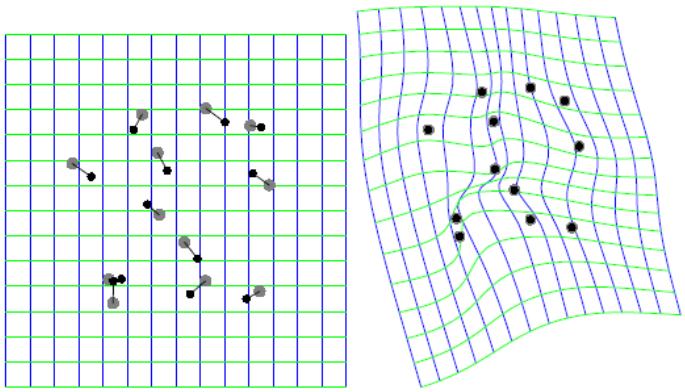


Figure 2.6: *Exemple d'interpolation basé sur les fonctions à base radiale de noyau plaque mince : l'interpolation passe exactement par les données utilisées pour calculer la fonction interpolante [95].*

2.4 Étalonnage supplémentaire pour la caméra dynamique

La méthode de calibrage que l'on vient de présenter permet de connaître la relation ζ liant une position dans I_s aux angles (α, β) de commande de la caméra dynamique tels que le projeté de P_w dans I_d soit confondu avec $\mathscr{C}_d$. On n'a donc pas de correspondance entre un pixel de la caméra statique et un pixel dans la caméra dynamique. Pour cela, il manque la connaissance de la relation entre un déplacement en pixel dans I_d et le déplacement angulaire correspondant, en fonction du zoom. Cette information permettra d'avoir la chaîne complète pour passer des coordonnées d'un point dans I_s à ces coordonnées dans I_d (illustrée par la flèche discontinue sur la figure 2.7).

De plus, lors du suivi, on souhaite adapter le zoom de la caméra dynamique à la taille de

la cible suivie. Il est donc nécessaire d'estimer la relation entre le paramètre de commande de zoom et le facteur d'échelle entre une distance dans I_s et la même dans I_d.

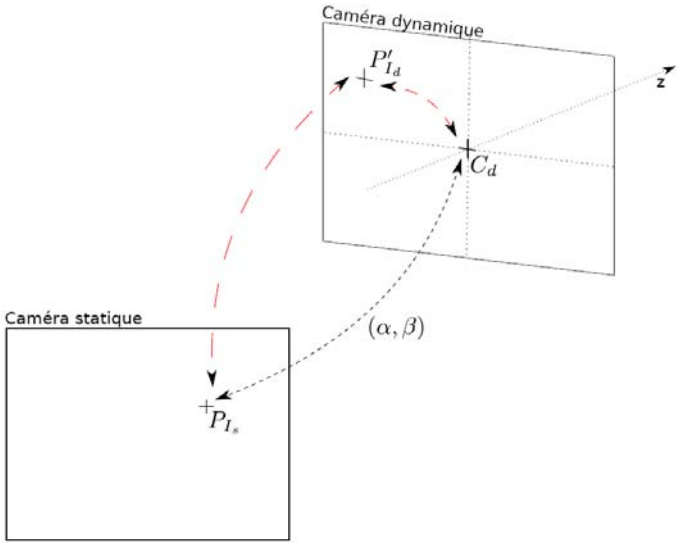


Figure 2.7: *Schéma présentant les relations entre la caméra statique et la caméra dynamique. La flèche en pointillée schématise le résultat de la méthode de calibrage qui donne une correspondance entre une position dans* I_s *et* $\mathscr{C}_d$ *pour un couple d'angle* (α, β) *donné. Les flèches discontinues illustrent les relations manquantes utiles pour mettre en place une méthode de suivi.*

Dans la suite de cette partie, on va exposer les procédures mises en place pour déterminer les relations suivantes :

1. relation entre un déplacement en pixel dans I_d et le déplacement angulaire correspondant en fonction du zoom,
2. relation entre le paramètre de commande de zoom et le facteur d'échelle entre une distance dans I_s et la même dans I_d.

Ces procédures sont spécifiques à la caméra dynamique utilisée lors de cette thèse pour les applications de suivi : c'est une caméra dôme de la marque AXIS.

2.4.1 Relation entre un déplacement en pixel et le déplacement angulaire correspondant en fonction du zoom

Le procédé utilisé pour estimer la relation est le suivant. On place la mire dans la scène. Pour un angle en site fixé β et pour différentes valeurs de zoom, on déplace la caméra dynamique N fois d'un même pas angulaire α' prédéfini. On refait les mêmes acquisitions mais pour un angle en azimut fixé α et pour différentes valeurs de zoom, on déplace la caméra dynamique N fois du même pas angulaire β' prédéfini.

On obtient les résultats illustrés sur la figure 2.8.

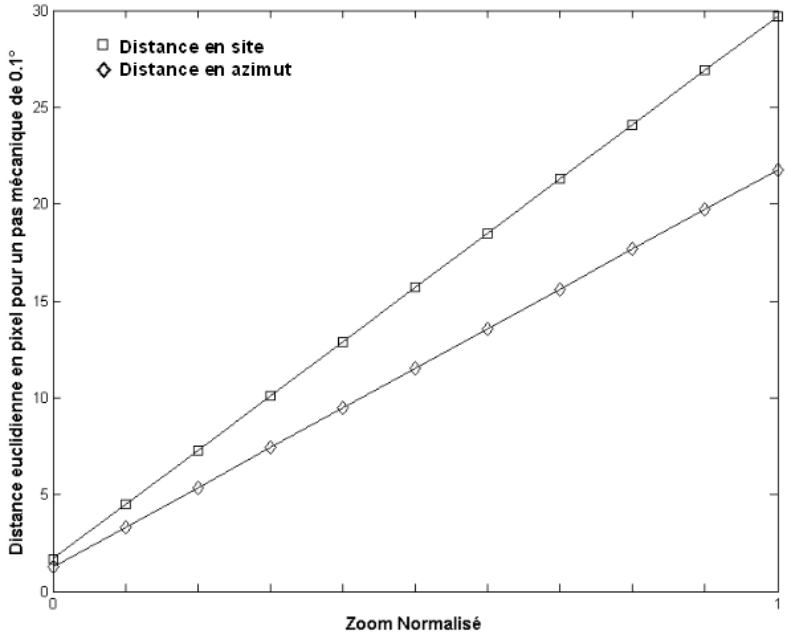


FIGURE 2.8: *Illustration de la relation linéaire liant un déplacement en pixel dans la caméra dynamique et le déplacement angulaire minimale de la caméra dynamique (*$0.1\,^\circ$ *dans notre cas) selon les deux axes de rotations.*

On constate que l'on peut approcher les résultats par une régression linéaire. Les relations linéaires, liant un déplacement en pixel Δd^{ref} dans I_d à un déplacement angulaire prédéfini (dans notre cas, le déplacement angulaire minimal de la caméra dynamique, c'est-à-dire $\Delta\alpha_{ref} = \Delta\beta_{ref} = 0.1°$) pour un zoom Z connu, sont données par les équations suivantes :

$$\Delta d_x^{ref} = a_\alpha Z + b_\alpha \tag{2.5}$$
$$\Delta d_y^{ref} = a_\beta Z + b_\beta \tag{2.6}$$

Sur une première approximation, on considère qu'il y a proportionnalité entre un déplacement en pixel Δd_x et un déplacement angulaire $\Delta\alpha$ et respectivement, Δd_y et $\Delta\beta$. On peut ainsi lier un déplacement en pixel à un déplacement angulaire de la caméra dynamique selon le zoom par une relation linéaire. On verra au chapitre 3 que la caméra utilisée lors des tests vérifie cette hypothèse. Les équations sont les suivantes pour un zoom Z connu :

$$\bullet\ \Delta\alpha(Z) = \frac{\Delta d_x\ \Delta\alpha_{ref}}{\Delta d_x^{ref}} \qquad \bullet\ \Delta d_x(Z) = \frac{\Delta\alpha\ \Delta d_x^{ref}}{\Delta\alpha_{ref}} \tag{2.7}$$

$$\bullet\ \Delta\beta(Z) = \frac{\Delta d_y\ \Delta\beta ref}{\Delta d_y^{ref}} \qquad \bullet\ \Delta d_y(Z) = \frac{\Delta\beta\ \Delta d_y^{ref}}{\Delta\beta ref} \tag{2.8}$$

2.4.2 Relation entre le paramètre de zoom et un facteur d'échelle

Soit un objet de dimension l pixels dans l'image de la caméra statique. Pour N valeurs de zoom, on va déterminer la taille L_{Z_n} en pixels de l'objet dans l'image de la caméra dynamique en fonction du zoom Z_n. Ensuite, on établit une table de correspondance liant le facteur d'échelle $\Gamma_{Z_n} = \frac{L_{Z_n}}{l}$ au zoom Z_n, illustrée sur la figure 2.9. On remarque que les N valeurs de Γ_{Z_n} peuvent s'approximer par une loi linéaire :

$$\Gamma(Z) = a_\Gamma Z + b_\Gamma. \tag{2.9}$$

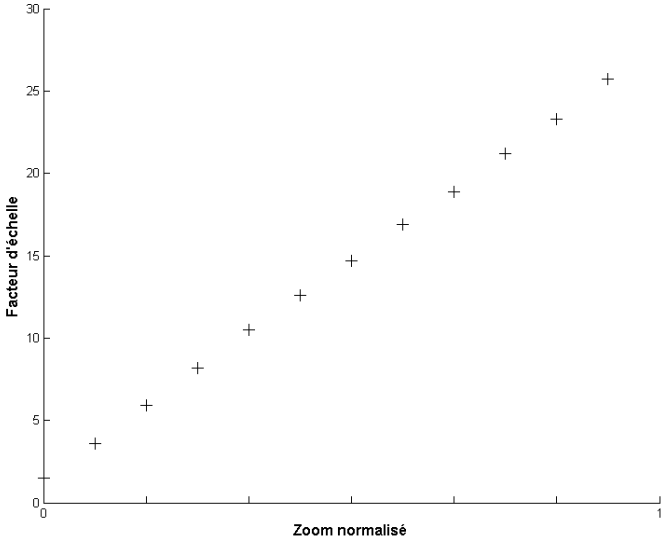


Figure 2.9: *Illustration de la relation linéaire liant le facteur d'échelle (rapport entre une distance dans la caméra dynamique et la même dans la caméra statique) et le zoom.*

2.5 Conclusion

Au final, grâce au calibrage de la paire de calibrage proposé et à l'étalonnage spécifique de la caméra dynamique, on a les relations suivantes :

◇ Pour tout couple (x_s, y_s) de I_s, on a

$$(\alpha, \beta) = \zeta(x, y),$$

◇ Pour tout couple (α, β) tel que le projeté de $\mathscr{C}_d$ dans la caméra statique appartienne à I_s, on a

$$(x, y) = \zeta^{-1}(\alpha, \beta),$$

◇ Pour tout couple (x_d, y_d) de I_d, pour une valeur de zoom Z connue, on a

$$\alpha_{x_d} = \alpha_{courant} + \Delta\alpha\,(Z, x_d - x_{\mathscr{C}_d}) \tag{2.10}$$

$$\beta_{y_d} = \beta_{courant} + \Delta\beta\,(Z, y_{\mathscr{C}_d} - y_d) \tag{2.11}$$

◇ Pour tout couple $(\Delta\alpha, \Delta\beta)$, pour une valeur de zoom Z connue, on a

$$x_d = x_{\mathscr{C}_d} + \Delta d_x\,(Z) \tag{2.12}$$

$$y_d = y_{\mathscr{C}_d} - \Delta d_y\,(Z) \tag{2.13}$$

Afin de schématiser les relations obtenues, on note Δ la fonction liant un couple de coordonnées de I_d à un couple d'angle (α, β) correspondant aux équations de 2.10 à 2.13. On peut donc transformer les coordonnées (x_s, y_s) d'un pixel de I_s en coordonnées (x_d, y_d) de I_d et inversement :

$$(x_s, y_s) \overset{\zeta}{\leftrightarrow} (\alpha, \beta) \overset{\Delta}{\leftrightarrow} (x_d, y_d)$$

Pour la suite du mémoire, on va noter $\mathrm{f}_{s\to d}$ la composition de transformations passant les coordonnées d'un point dans I_s aux coordonnées de ce point dans I_d.

Chapitre 3

Résultats et discussion de la méthode de calibrage

La méthode de calibrage proposée permet de faire le lien entre les coordonnées (x, y) du point P_{I_s} de I_s et les angles (α, β) tels que P_{I_d} soit confondu avec $\mathscr{C}_d$. Afin d'estimer la précision de notre méthode, on évalue l'erreur commise entre la position réelle de P_w dans I_d et la position recherchée, c'est-à-dire $\mathscr{C}_d$. Pour cela, il est nécessaire d'avoir une vérité terrain. Ayant fait le choix de ne pas faire de calibrage fort de la paire de caméras, les coordonnées 3D d'un point de la scène ne sont pas accessibles. Pour remédier à cela, on propose de mesurer l'erreur de position dans les repères caméras en utilisant une mire afin d'estimer avec précision la position des coordonnées des points caractéristiques.

On a choisi d'utiliser une mire composée d'une ellipse noire sur un fond blanc qui soit visible par les deux caméras, illustrée sur la figure 3.1. Le point caractéristique de cette mire est le centre de gravité de l'ellipse noire. Afin de déterminer avec précision les coordonnées de son centre, on utilise la méthode de seuillage d'Otsu [77]. Cela consiste à séparer la région d'intérêt en deux classes. Dans notre cas, la première classe correspond aux pixels clairs qui seront labellisés comme pixels blancs et la seconde aux pixels sombres labellisés comme pixels noirs. Ensuite, il reste à estimer avec une précision subpixélique les coordonnées du centre de gravité des pixels noirs.

Les résultats obtenus seront ramenés à une valeur angulaire afin de comparer à la précision mécanique des caméras. Le déplacement angulaire minimal (pas mécanique) des caméras utilisées lors de cette thèse est de l'ordre de 0.1°.

3.1 Calibrage du système de vision : méthode analytique ou méthode par apprentissage

Ayant choisi de ne pas faire de calibrage fort et ne connaissant pas de façon précise la géométrie du système, les paramètres angulaires de la caméra dynamique ne peuvent

Figure 3.1: *Illustration de la cible elliptique utilisée pour évaluer la précision de la méthode de calibrage : à gauche dans la caméra statique et à droite dans la caméra dynamique.*

s'estimer que de deux manières : soit en modélisant de façon simple le système (solution analytique), soit par la méthode de calibrage basée sur l'apprentissage, comme celle que l'on propose.

Supposons que le mécanisme de la caméra dynamique vérifie les hypothèses suivantes : son centre optique est confondu avec son centre de rotation et les axes de rotation sont alignés avec la géométrie du capteur CCD (figure 3.2). Les deux caméras étant quasiment alignées et très proches, on peut faire l'hypothèse que leurs centres optiques sont confondus. Dans ces conditions, la transformation permettant de passer des coordonnées d'un point de I_s aux angles nécessaires pour centrer la caméra dynamique sur ce point est simple. En effet, la transformation exprimant le passage du repère de la caméra statique à la caméra dynamique revient à une composition de rotations. La relation liant les coordonnées d'un point dans I_d aux paramètres angulaires à appliquer pour centrer la caméra dynamique sur ce point peut s'estimer simplement de manière analytique.

On note (x_p, y_p) les coordonnées du centre de la cible dans I_d, k_u et k_v les facteurs d'échelle permettant de faire la conversion d'unité entre le monde réel et l'image et f la focale de la caméra. Les paramètres angulaires (α, β), tels que la caméra dynamique pointe sur la cible, sont donnés par les relations suivantes :

$$\alpha = arctan\left(\frac{x_p - x_{\mathscr{C}_d}}{k_u f}\right) \tag{3.1}$$

$$\beta = arctan\left(\frac{y_p - y_{\mathscr{C}_d}}{k_v f}\right) \tag{3.2}$$

On constate que, dans cette configuration, l'estimation des angles de rotation ne dépend que d'une des coordonnées du point caractérisant la cible : $\alpha = g(x_p)$ et $\beta = g(y_p)$. Ainsi, si l'on fait varier α en gardant l'angle β fixe, on va obtenir une trajectoire apparente d'un point de I_d rectiligne, de même si on fait varier l'angle β et que α reste fixe.

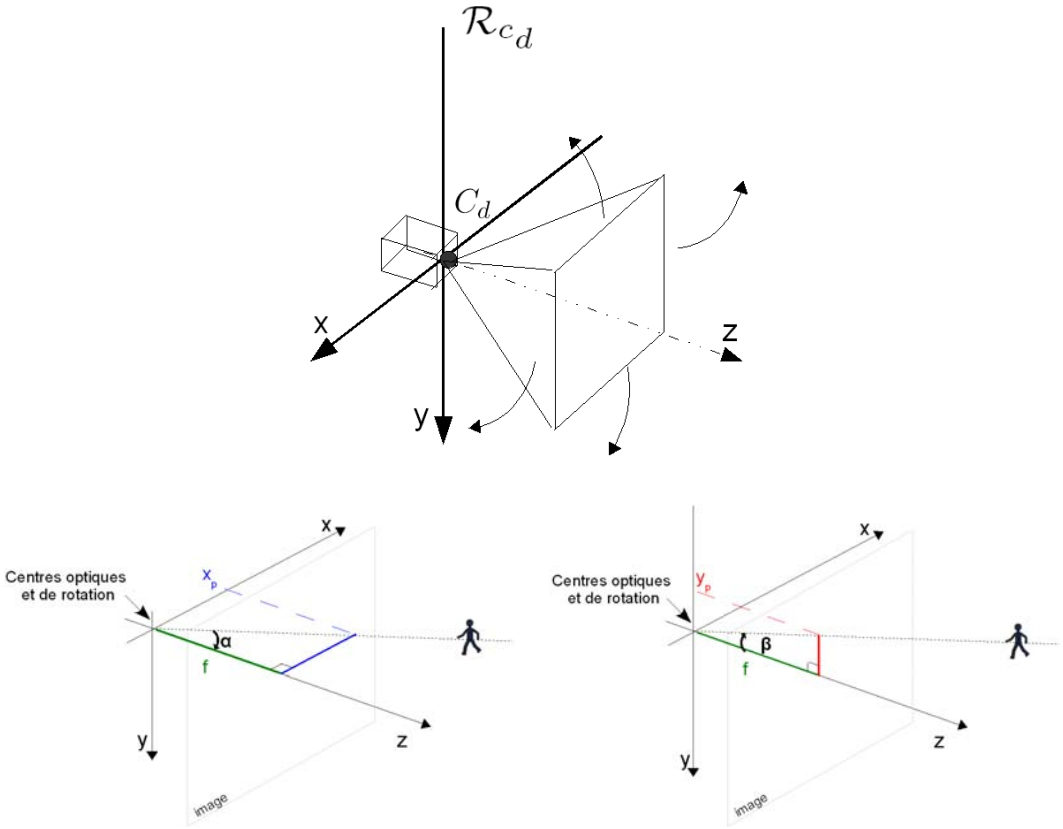


FIGURE 3.2: *Illustration du calcul analytique pour estimer les paramètres angulaires de commande (α, β) dans le cas où le centre optique est confondu avec le centre de rotation.*

3.1.1 Discussion sur une modélisation simple de la rotation

En pratique, les mécanismes réels de caméras dynamiques vérifient rarement les hypothèses énoncées précédemment. En effet, il est difficile pour des produits industriels que le centre optique soit correctement juxtaposé avec les deux axes de rotation. Cela implique que le mouvement apparent d'un point dans l'image ne suivra plus une trajectoire rectiligne mais une trajectoire quelconque dans l'image, et ceci, de façon décorrélée de toute distorsion optique. Afin d'illustrer ces propos, on a testé les deux caméras dynamiques que nous avions à notre disposition : la caméra Pan-Tilt-Zoom AXIS 213 (notée PTZ par la suite) et la caméra dôme AXIS 233D.

Les données ont été obtenues de la manière suivante. La mire est placée dans la scène observée par les caméras dynamiques. La caméra est centrée sur la mire pour un couple $(\alpha_{ref}, \beta_{ref})$ donné. Ensuite, on acquiert une séquence d'image en faisant varier la valeur α de l'angle azimut avec β_{ref} fixe et inversement. Enfin, pour chaque couple (α, β), on détermine les coordonnées du centre de la cible dans l'image liée à cette position.

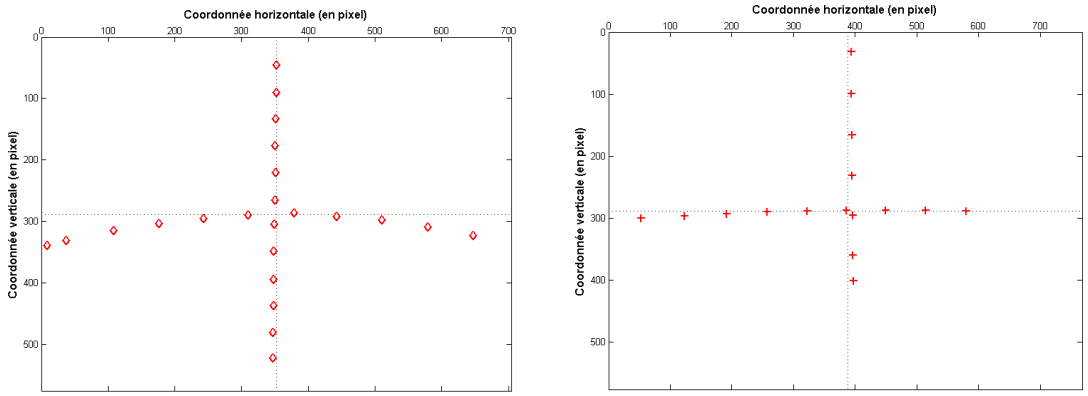


Figure 3.3: *Illustration de la trajectoire apparente d'un point dans l'image lors d'une rotation en azimut et en site de la caméra dynamique : à gauche les résultats pour la caméra Pan-Tilt-Zoom AXIS 213 et à droite pour la caméra dôme AXIS 233D.*

On obtient, ainsi, les résultats illustrés sur la figure 3.3. On constate que le mouvement du centre de la mire décrit une courbe selon l'axe horizontal lorsque la mire est observée par la caméra PTZ alors que, dans le cas de la caméra dôme, le mouvement est quasiment rectiligne. L'angle de rotation α de la caméra PTZ n'est pas seulement lié à l'abscisse du centre de la cible : on a $\alpha = g(x_p, y_p)$. Par conséquent, la méthode analytique pour évaluer les paramètres angulaires ne peut pas s'appliquer pour cette caméra. Il faut une modélisation plus générale. Par contre, au vu des courbes de la figure 3.3, le mécanisme de la caméra dôme semble très proche de la modélisation idéale d'un mécanisme de rotation d'une caméra dynamique. On peut déterminer, avec une erreur limitée, les paramètres angulaires de la caméra avec la méthode analytique simple proposée.

Au vu de la précision du mécanisme de rotation de la caméra dôme, les relations linéaires liant un déplacement en pixel à un déplacement angulaire de la caméra dynamique proposée

au chapitre précédent (équations 2.7 et 2.8) suffissent pour la caméra dôme qui sera utilisée dans la suite des expérimentations.

3.1.2 Discussion sur l'influence du zoom

La focale d'une caméra dynamique est fonction du zoom de celle-ci. Ainsi, même si le mécanisme de rotation est optimal à un zoom donné, une variation du zoom, c'est-a-dire un changement au niveau des lentilles composant le système optique de la caméra dynamique, peut entraîner une déviation de la position du centre optique : ceci a été évalué par Gardel lors de sa thèse [35]. On a pu mettre en évidence cette déviation lors d'expérimentations.

Les données ont été obtenues de la manière suivante. On positionne la caméra dynamique telle que le centre de I_d coïncide avec le centre de la mire. Ensuite, on fait varier la valeur du zoom. On obtient ainsi une séquence d'image pour une position fixe et différentes valeurs du zoom. Pour chaque image de la séquence, on détermine les coordonnées du centre de la mire. Ensuite, on estime la distance euclidienne entre les coordonnées du centre de la mire au zoom initial et ceux des zooms plus élevés, ramené à une erreur en degré.

On obtient les résultats présentés sur la figure 3.4. On note que dans le cas de la caméra PTZ, la déviation est croissante en fonction du zoom. Dans le cas de la caméra dôme, la mesure d'angle obtenue s'apparente à la précision mécanique et ce quelque soit le zoom : l'optique est donc de très bonne qualité. Ceci implique, dans le cas de la caméra PTZ, que la déviation du centre optique due à un changement du zoom n'est pas négligeable et que l'on ne peut pas mettre en pratique l'approche analytique pour déterminer la commande de la caméra.

Afin d'avoir une méthode générique, c'est-à-dire applicable à n'importe quel type de caméra dynamique, on a choisi de mettre en place une méthode de calibrage faible qui va implicitement prendre en compte par apprentissage les imperfections du mécanisme de rotation de la caméra mais aussi celles des lentilles (déviation du centre optique et distorsions dans l'image).

3.2 Approche manuelle ou automatique

Parmi les méthodes de calibrage faible citées précédemment (paragraphe 1.2), notre méthode se rapproche le plus de celle proposée par Zhou *et al.* Leurs méthodes reposent sur un établissement d'un nombre fini de correspondances position/angles. Le choix des points à apprendre est fait par l'opérateur. La connaissance des correspondances est manuelle : l'opérateur centre à la main la caméra dynamique sur la position prédéfinie. Ceci implique une variance intra et inter opérateur. En effet, d'un opérateur à l'autre et d'une fois sur l'autre, pour une même scène, le choix des points ne sera pas identique ainsi que la précision de l'apprentissage. Pour une même scène, cette solution ne donne pas l'assurance de la reproductibilité des résultats à l'inverse de la méthode automatique de calibrage que l'on propose. Par contre, la méthode de Zhou *et al.* assure une meilleure fiabilité des résultats :

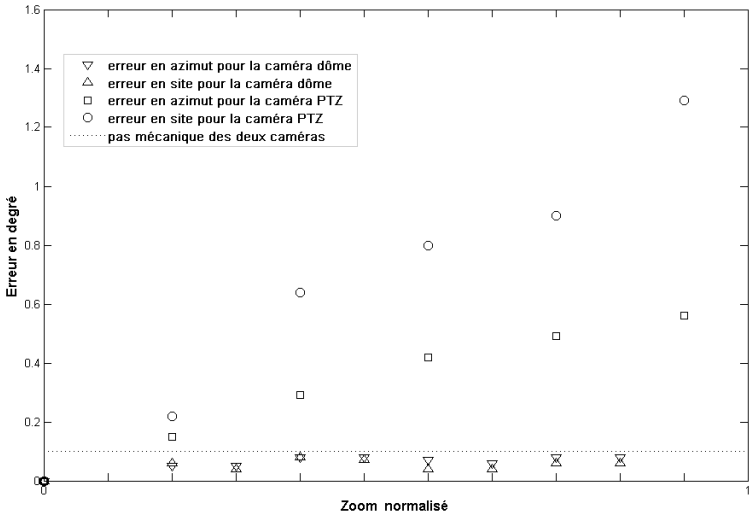


Figure 3.4: *Illustration de la déviation du centre optique en fonction du zoom pour les deux caméras dynamiques. L'erreur est exprimée en degré afin de la comparer avec la précision mécanique de chaque caméra.*

une personne sera toujours plus performante qu'un traitement logiciel automatique qui a ses limites comme par exemple la mise en correspondance de points dans des zones homogènes.

De plus, comme notre méthode de calibrage est automatique, il est beaucoup plus aisé de mettre en place une mise à jour régulière de la méthode que de faire déplacer spécialement un opérateur sur le site. En effet, pour une scène pour laquelle à certains horaires de la journée, on sait qu'il n'y a personne, il est facile de programmer l'exécution de notre méthode de calibrage.

3.3 Résultats et précision de la méthode

Notre méthode de calibrage fait la distinction entre les points appris lors de la première étape du calibrage et les points interpolés suite à la deuxième étape du calibrage. Les points appris servent de base à la fonction d'interpolation. On va présenter des résultats d'expérimentations illustrant, tout d'abord, la précision obtenue suite à l'apprentissage des correspondances position/angles, puis, en prenant en compte l'étape d'interpolation, la précision de la commande obtenue sur des points quelconques de la scène.

3.3.1 Précision de l'apprentissage

Afin d'évaluer la précision de l'asservissement visuel lors du calibrage, on va positionner la mire sur un certain nombre de points du quadrillage tel que le centre de gravité de la mire soit confondu, au pixel près, avec le point du quadrillage. Ensuite, on applique la correspondance position/angles apprise lors du calibrage afin de centrer la caméra dynamique sur cette position. Les coordonnées du centre de gravité de la mire dans I_d sont déterminés. Enfin, on estime l'erreur entre la position du centre de gravité de la mire et le centre de I_d. Cette erreur est ramenée à une erreur angulaire.

Résultats avec la caméra dôme

Comme on a pu le voir précédemment, la déviation due au zoom dans le cas de la caméra dôme est de l'ordre de la précision mécanique de la caméra. Il n'est donc pas nécessaire d'appliquer l'étape de calibrage pour les différents zooms. On va donc seulement présenter des résultats de la méthode de calibrage au zoom initial pour cette caméra dynamique.

Les points du quadrillage, sur lesquels on va s'appuyer pour la discussion, sont les points cerclés de noir sur la figure 3.5. Les numéros correspondent à leurs emplacements dans la liste des points constituant le quadrillage (quadrillage initial de 124 points). Les points 3, 7 et 13 sont des points dont la densité de points SIFT extraits dans le voisinage du point est importante. Les points 43 et 61 ont plutôt une densité moyenne. Le point 93 a une densité très faible.

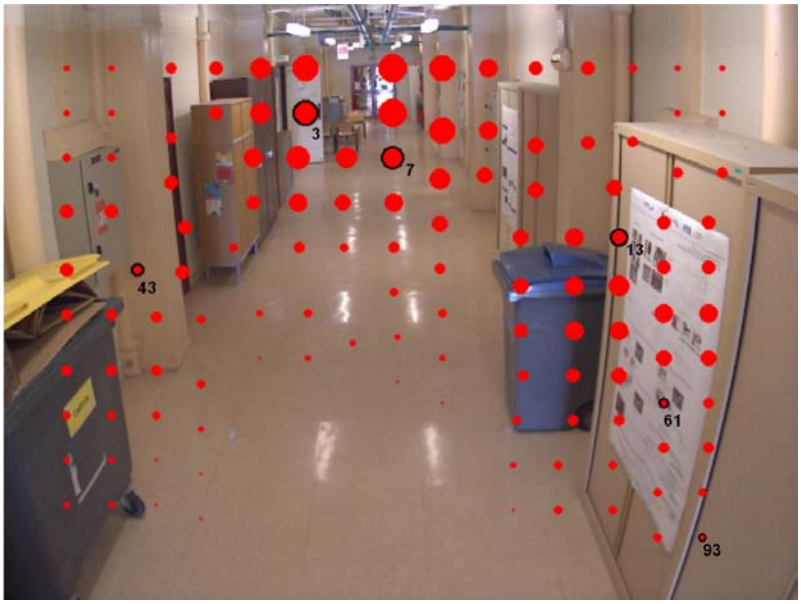


Figure 3.5: *Quadrillage d'apprentissage lors du calibrage. Les points cerclés de noirs servent de base à la discussion sur la précision du calibrage. La taille des cercles représentent la densité de points SIFT extraits dans leurs voisinage.*

Figure 3.6: *Voisinage dans* I_s *des six points du quadrillage cerclés de noir sur la figure 3.5.*

On remarque sur la figure 3.6 que le voisinage des points de label 3 et 7 englobe une région de la scène présentant de grandes variations dans la géométrie 3D : une profondeur de plusieurs dizaines de mètres. Le voisinage des points de label 13 et 43 présente une plus grande homogénéité de la géométrie 3D de la scène : une profondeur de quelques mètres. Le voisinage des points de label 61 et 93 est composé principalement d'un seul plan.

Les résultats des expérimentations sont présentés dans le tableau 3.1.

	Point 3	Point 7	Point 13	Point 43	Point 61	Point 93
Erreur en degré en azimut	0.54°	0.13°	0.13°	0.28°	0.09°	0.6°
Erreur en degré en site	0.02°	0.3°	0.16°	0.29°	0.08°	0.32 °

Table 3.1: *Tableau donnant l'erreur commise lors du calibrage sur l'apprentissage de la correspondance position/angles.*

La base de notre méthode de calibrage est basée sur la mise en correspondance de points d'intérêt qui servent à estimer la projection des coordonnées du point du quadrillage dans I_d. L'estimation de la projection sera d'autant plus robuste que le nombre de bonnes correspondances est important. Le nombre de correspondance dépendant du nombre de points SIFT extraits dans le voisinage des points du quadrillage, on peut supposer que la précision des premiers points du quadrillage sera meilleure que celle des points de fin de liste. Le voisinage des points 13 et 43 et des points 61 et 93 est assez semblable (figure 3.6). La différence notable entre les deux cas est la densité des points extraits par la méthode SIFT de Lowe dans la zone d'intérêt. On constate que la précision des points 43 et 93 est

moins bonne que celle des points 13 et 61 bien que l'environnement soit semblable. Ainsi, l'hypothèse selon laquelle la précision est d'autant meilleure que le point du quadrillage à un voisinage ayant une grande densité de points SIFT se trouve illustrée par ces résultats. C'est pourquoi nous avons choisi de construire un quadrillage en fonction de l'information extraite de la scène 3D et non d'utiliser un quadrillage régulier.

Par contre, la précision obtenue pour les points 3 et 7 est bien plus mauvaise que celle obtenue pour le point 61 alors que les points 3 et 7 sont en début de liste. On remarque que le voisinage de points 3 et 7 représente une région de la scène présentant de fortes disparités au niveau de sa géométrie 3D alors que celui du point 61 est plan. Or, la transformation utilisée entre les deux capteurs est une homographie. Cela suppose que les deux images comparées sont planes ce qui n'est pas le cas pour les points 3 et 7 ce qui dégrade la précision du calcul au contraire du point 61. Le couloir est un cas extrême : scène étroite et toute en profondeur. Il est donc difficile, dans ces conditions, que le voisinage d'un point se situe sur un même plan, à la différence d'un bureau par exemple ou d'une scène en extérieure comme on peut le voir sur la figure 3.7. Le choix de l'homographie n'est donc pas optimale si l'on souhaite une précision maximale avec ce type de calibrage. On peut envisager, après la convergence de la méthode actuelle, de rajouter une étape de raffinement en restreignant la zone d'intérêt autour du point pour approcher au mieux un plan et de mesurer la corrélation entre les deux zones d'intérêt. Pour notre application de suivi, la précision obtenue avec la méthode de calibrage basée sur un calcul d'homographie est suffisante.

Figure 3.7: *Scènes présentant plus de plans que dans le cas d'un couloir.*

Résultats avec la caméra PTZ

Comme on a pu le voir précédemment, la déviation due au zoom dans le cas de la caméra PTZ est très importante. Ceci a pour conséquence que dans les forts zooms, l'objet d'intérêt ne soit plus visible en entier. Pour ce type de matériel, il est donc nécessaire d'appliquer l'étape de calibrage pour les différents zooms. La figure 3.8 présente l'erreur moyenne commise pour différents zooms pour un ensemble de points du quadrillage obtenus

pour la caméra PTZ.

Sur chaque graphe, on a reporté l'erreur moyenne commise pour différentes valeurs de zoom ainsi que la déviation observée d'un point lorsque l'on zoome. On constate que l'erreur commise est moindre avec le calibrage : de l'ordre de $0.2° - 0.3°$ au lieu d'erreur pouvant aller jusqu'à $0.6° - 0.8°$. Ce calibrage permet donc de corriger les faiblesses du mécanismes de la caméra PTZ. Ainsi, lors d'une application de suivi, ce calibrage va permettre de pouvoir garder en continue une cible visible dans son intégralité au zoom voulu.

3.3.2 Précision de l'interpolation

Maintenant, on va s'intéresser à la précision globale de notre système c'est-à-dire en prenant en compte l'étape d'interpolation. On a choisi un certain nombre de points répartis dans la scène en dehors des points appris (figure 3.9). Les points de label a, b et c sont positionnés au milieu de point appris. Les points de label d et e sont placés dans une zone où il y a très peu de points appris à cause de son homogénéité (sol du couloir).

Les résultats des expérimentations sont présentés dans le tableau 3.2.

	Point a	Point b	Point c	Point d	Point e
Erreur en degré en azimut	0.12°	0.26°	0.13°	0.0.7°	0.5°
Erreur en degré en site	0.11°	0.02°	0.11°	0.34°	0.5°

Table 3.2: *Tableau donnant l'erreur commise pour des points qui n'ont pas été appris lors de la première phase de calibrage.*

On constate que la précision obtenue avec l'interpolation est du même ordre de grandeur que celle pour les points appris. De plus, il n'y a pas de grande disparité dans les précisions : quelque soit la position du point dans I_s, l'erreur commise est assez homogène.

Le point a est celui qui a la meilleure précision. Outre le fait que ce point fait parti des points appris, le voisinage de a est un plan. La précision des points b et c est moins bonne. A la différence du point a, le voisinage de ces deux points n'est pas un plan. Comme on a pu le voir précédemment, cela influence la précision obtenue pour les points appris. Cette influence se retrouve aussi au niveau de l'interpolation. Comme on peut s'y attendre, la précision obtenue pour des points hors de la densité des points du quadrillage, comme les points d et e est moindre mais reste tout a fait suffisant pour notre application.

Les méthodes d'interpolation sont très sensibles aux données servant de base au calcul d'interpolation. Plus les données sont réparties de manière dense dans tout l'espace, et plus la méthode d'interpolation est précise. Ainsi, si l'on souhaite améliorer nos résultats, il faut que les points du quadrillage soient en nombre suffisants et bien répartis. Pour cela, on peut choisir, lors de la construction du quadrillage, de diminuer la distance entre deux points afin d'augmenter le nombre de points et de rendre plus dense la quadrillage. Une autre solution est d'agrémenter la scène de manière la plus naturelle possible pour amener

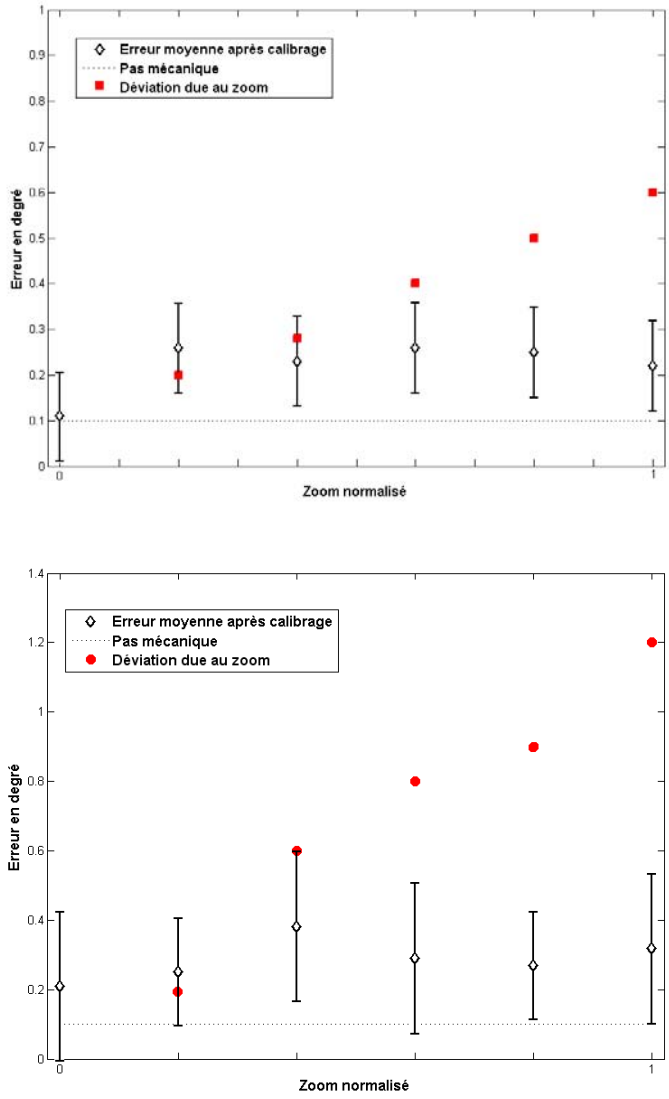


FIGURE 3.8: *Graphes représentant la précision de la méthode de calibrage à différents zooms dans le cas de la caméra PTZ : en haut, les résultats en azimut et en bas les résultats en site.*

Figure 3.9: *Les croix représentent les points interpolés pour lesquels on va estimer la précision globale de notre calibrage. Les points représentent les point du quadrillage qui ont pu être appris lors de la première étape du calibrage.*

de l'information là où il en manque : la figure 3.10 illustre cette possibilité. Dans certains cas, le manque d'information est inhérent à la scène : par exemple, pour une scène en extérieur (image de gauche de la figure 3.7), on ne pourra pas y remédier comme pour une scène en intérieur.

3.4 Conclusion

Dans cette première partie, nous avons proposé une nouvelle solution de modélisation optique et géométrique du système de vision hybride qui tend vers un dispositif automatique et autonome. Cette modélisation donne accès directement à une relation entre un déplacement en pixel dans une image et les paramètres angulaires à appliquer à la caméra dynamique pour que celle-ci se centre sur la position d'intérêt.

Une application immédiate de notre méthode de calibrage est d'utiliser la paire de caméras comme un système maître-esclave. Cela consiste à désigner une position dans la caméra statique et d'amener automatiquement la caméra dynamique sur cette zone d'intérêt afin d'obtenir une meilleure résolution, ponctuelle, de la scène. La figure 3.11 illustre les images que l'on peut obtenir avec ce système pour des environnements intérieurs et extérieurs. On constate que la définition et la précision est suffisante pour envisager l'ajout d'un module de reconnaissance de geste et/ou d'identification de personne.

Maintenant que l'on sait commander la caméra dynamique ponctuellement et à zoom constant, on peut envisager de mettre en place un système de suivi dynamique tel que la

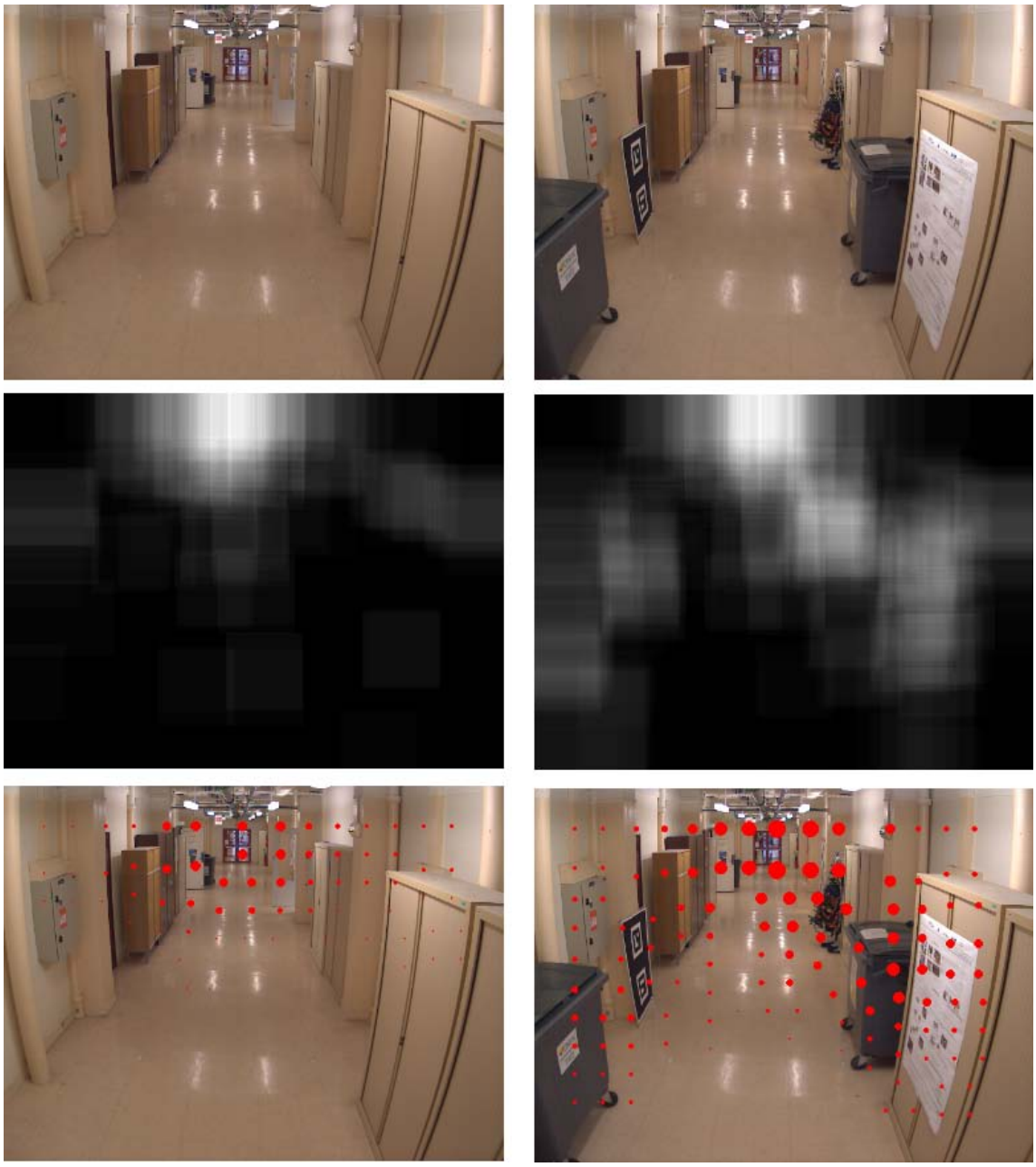


FIGURE 3.10: *La première ligne représente la scène observée par la caméra statique : un couloir. La ligne du milieu représente l'estimation non paramétrique (méthode des fenêtres de Parzen) de la densité de probabilité des points d'intérêt dans* I_s*. La ligne du bas montre la grille extraite de* I_s*, plus les cercles sont importants et plus la densité de probabilité est importante.*

Figure 3.11: *Exemples d'images résultant de l'application directe de la méthode de calibrage en intérieur (première ligne) et en extérieur (dernière ligne). La croix dans les images de la caméra statique (colonne de gauche) représente le position sur laquelle on veut centrer la caméra dynamique. Les images obtenues avec la caméra dynamique (colonne de droite) permet d'envisager des applications de reconnaissance.*

caméra dynamique suive en continu la personne en adaptant son zoom à la taille de la cible.

Deuxième partie

Suivi multi-caméras de personne

Chapitre 4

Méthodes de suivi : Etat de l'art

Classiquement, les méthodes de suivi d'objet dans une séquence vidéo s'articulent autour de diverses étapes : tout d'abord l'initialisation de l'élément recherché (cible) par la définition d'un modèle de référence, puis la recherche automatique de la cible dans les images suivantes du flux vidéo en cherchant l'élément candidat ayant la plus grande ressemblance avec le modèle de référence. Cette recherche peut s'interpréter comme la minimisation d'une distance entre le modèle et la cible candidate représentant la dissimilarité entre le modèle et la cible. La notion de distance peut être caractérisée au sens de la cohérence spatiale de la trajectoire de l'objet et de la similarité d'apparence visuelle entre l'observation et le modèle.

La contrainte spatiale de trajectoire s'exprime à travers la définition d'un modèle dynamique a priori qui sert à orienter le suivi selon un type de mouvement donné. Les véhicules se prêtent par exemple assez bien à une modélisation dynamique, alors que les trajectoires des personnes sont beaucoup plus difficiles à appréhender du fait de leur caractère plus « aléatoire »(nombre important de degrés de liberté).

Dans le cas du suivi basé sur l'apparence visuelle, la définition du modèle d'apparence est critique dans la mesure où elle conditionne le type d'information que l'on doit rechercher dans l'image. Ce modèle doit décrire avec suffisamment de précision la cible pour la discriminer des autres objets et du fond. En même temps, il doit également être robuste aux variations d'apparence survenant lors du mouvement de la cible, ou dans des cas de changements de point de vue ou d'illumination de la scène. Typiquement, le suivi de personnes est un cas particulièrement complexe du fait de la multitude d'attitudes et de positions qu'elles peuvent prendre. La variabilité d'apparence des objets nécessite dans certains cas une adaptation du modèle au cours du temps. Cette mise à jour soulève une problématique extrêmement délicate puisqu'il n'est en général pas possible de déterminer de manière robuste le critère de décision de cette mise à jour.

Par ailleurs, la méthode de suivi doit pouvoir gérer les occultations partielles ou les occultations totales temporaires.

Le suivi d'un objet dans une séquence vidéo s'articule autour de trois choix essentiels : une représentation de l'objet adaptée et suffisamment discriminante pour retrouver la cible

dans la scène, une méthode de détection de l'objet dans la scène et une méthode de suivi. Certains algorithmes combinent les approches de détection et de suivi d'objet.

Ces différents points vont être abordés dans un état de l'art dédié aux méthodes pour du suivi mono-caméra. La gestion de suivi de cible dans un contexte multi-caméras sera traitée dans une deuxième partie.

4.1 Méthodes de suivi d'objet mono-caméra

L'état de l'art proposé sur les méthodes de suivi dans le cas mono-caméra est limité aux méthodes applicables à tout type d'objet. On n'abordera donc pas les méthodes qui se basent sur des caractéristiques spécifiques à l'objet suivi comme des modèles d'objet articulé adaptés au suivi de personne (figure 4.1(e), 4.1(f)). Cet état de l'art est basé sur l'étude effectuée par Yilmaz *et al.* [111].

Dans un premier temps, on va présenter les diverses possibilités de représentation d'un objet ainsi que les différentes primitives utilisées dans les algorithmes de suivi. Puis, on détaillera les méthodes de détections et de suivi d'objets.

4.1.1 Représentation d'un objet et primitives visuelles associées

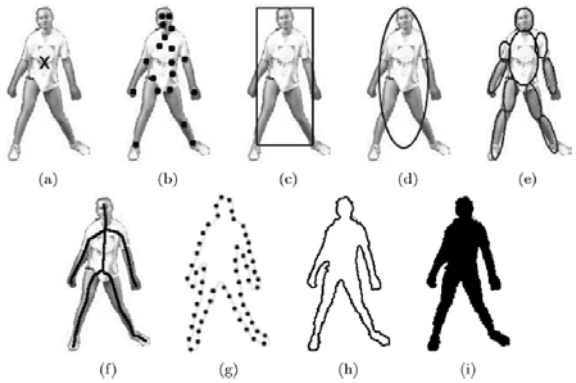


Figure 4.1: *Illustration des différents types de représentation d'un objet [111] : (a) centre de gravité, (b) ensemble de points, (c) rectangle englobant l'objet, (d) ellipse englobant l'objet, (e) et (f) modèles articulés, (g) points de contrôle sur le contour d'un objet, (h) contour de l'objet , (i) silhouette d'un objet.*

Un objet peut être représenté par sa forme et son apparence. Les différents types de représentation de la forme d'un objet usuellement utilisés pour le suivi d'objet sont les suivants :

◇ *Points*

Un objet peut être représenté par un point, classiquement son centre de gravité (figure 4.1(a)), ou par un ensemble de points (figure 4.1(b)) [104]. En général, la représentation d'un objet par un point est approprié au suivi d'objet de petite taille dans une image.

◇ *Formes géométriques simples*
La forme d'un objet peut être représentée par une forme géométrique simple (figure 4.1(c), 4.1(d)) comme un rectangle ou une ellipse [26]. Avec ce type de représentation, le mouvement de l'objet est habituellement modélisé par une translation, une transformation affine ou projective (homographie). Bien que cette modélisation soit plus appropriée pour représenter des objets rigides simples, elle est aussi utilisée pour le suivi d'objets non rigides.

◇ *Contour d'un objet*
L'objet peut être représenté par un contour qui définit les frontières de l'objet [110]. Ce type de représentation est utilisé dans le cas de suivi d'objets non rigides complexes.

L'objet peut aussi être caractérisé par son apparence, c'est-à-dire par un vecteur contenant des primitives visuelles issues des images le caractérisant. Les primitives usuellement utilisées sont :

◇ *La couleur* : L'espace colorimétrique RGB est l'espace le plus utilisé en traitement d'image. Ce type de primitive est très sensible au changement d'illumination de la scène et aux propriétés de réflexion des objets. D'autres systèmes permettent de représenter la couleur tels que le système XYZ obtenu par transformation linéaire de l'espace RGB ou l'espace de couleur représenté par la teinte et la luminosité (appelé HSV en anglais), le plus intuitif par rapport à la vision humaine. Cet espace présente l'avantage d'être invariant aux conditions d'illumination à condition que la saturation et la luminosité ne soient pas trop faibles [47].

◇ *Le gradient* : Comparé aux primitives couleurs, les primitives liés aux bords (*edge* en anglais) sont moins sensibles aux changements d'illuminations. Ce type de primitive est généralement associé à une représentation de l'objet par ses contours. Une étude proposant une évaluation d'algorithmes de détection de bords est proposé par Bowyer *et al.* [15].

◇ *Le mouvement* : La technique du flot optique représente un champ dense de vecteurs de déplacement caractérisant la translation de chaque pixel dans une région de l'image. C'est pourquoi cette primitive est généralement associée à des méthodes de segmentation et de suivi basé sur le mouvement. Les différentes approches liées au flot optique sont décrites plus en détail dans les publications de Barron [4], Beauchemin *et al.* [6] et Mitiche *et al.* [71].

◇ *La texture* : C'est une mesure de la variation d'intensité d'une surface caractérisant ses propriétés telles que sa régularité et sa granularité ce qui nécessite la définition d'un descripteur. De même que les bords, la texture est peu sensible au changement d'illumination. De nombreux descripteurs de texture existent : entre autre, les matrices de cooccurrence de niveau de gris de Haralick [40] calculées sur la base de l'énergie, l'entropie, l'homogénéité, la corrélation et la variance et les méthodes basées sur les ondelettes [66].

Parmi toutes ces primitives, la primitive couleur est celle qui est le plus souvent utilisée dans les algorithmes de suivi.

L'apparence peut être représentée par l'estimation d'une densité de probabilité : paramétrique sous la forme d'une gaussienne ou d'une mixture de gaussiennes ou non-

paramétrique, basée sur les fenêtres de Parzen ou un histogramme. Les primitives utilisées pour estimer la densité de probabilité peuvent être calculées à partir de régions spécifiques de l'image définies par la forme de l'objet (intérieur d'une région délimitée par une ellipse ou les contours de l'objet).

Les modèles actifs d'apparence modélisent à la fois la forme et l'apparence de l'objet. Généralement, la forme est définie par un jeu de points caractéristiques se situant sur le contour de l'objet et à l'intérieur : dans le cas d'un visage, les points pourront se situer sur le pourtour du visage, les yeux et la bouche. Pour chacun de ces points, un vecteur d'apparence lui est associé se basant sur la couleur, la texture ou l'amplitude du gradient. Ce type de représentation nécessite une phase d'apprentissage sur une base de données caractérisant l'objet.

Un objet peut aussi être modélisé par plusieurs vues. Une approche pour représenter ces différentes vues consiste à générer un sous-espace à partir d'une vue donnée. Ceci peut être fait avec une analyse en composante principale ou une analyse en composante indépendante. Une autre approche consiste à apprendre les différentes vues de l'objet grâce à une méthode de classification [78].

4.1.2 Détection d'objets

Une approche courante pour la détection d'objet est d'utiliser l'information venant d'une seule image : approche statique. Cependant, certaines méthodes se basent sur une information temporelle calculée à partir d'une séquence d'images afin de réduire les fausses détections. On parle dans ce cas d'approche dynamique. Les méthodes les plus populaires de détection d'objet peuvent se classer en quatre catégories : détection de points d'intérêt dans une image, détection de zones homogènes par une segmentation de l'image, détection de zones en mouvement par une modélisation du fond ou détection utilisant des méthodes de classification.

Détection de points dans une image

Les méthodes de détection de points dans l'image listent des points caractéristiques de l'image ayant une texture riche dans le voisinage de ces points. Ce type de méthodes est depuis longtemps utilisé dans le contexte du mouvement, de la stéréo et du suivi. La qualité d'un point d'intérêt dépend notamment de son invariance aux changements d'illumination. Dans la littérature, les méthodes les plus usuelles de détection de points d'intérêt sont le détecteur de point de Harris [41], le détecteur KLT [93] et le détecteur SIFT [63]. La figure 4.2 montrent les résultats de détection obtenus avec ces trois méthodes. Pour une étude comparative des méthodes de détections de points d'intérêt, on peut se reporter à l'étude de Mikolajczyk et Schmid [70]. Une fois la détection effectuée, une étape de mise en correspondance est effectuée afin d'apparier les points d'une image à l'autre dans un but de suivi d'objet.

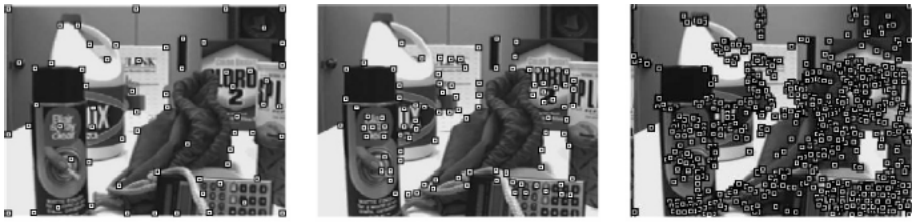


FIGURE 4.2: *Illustration de détecteur de points d'intérêt [111] : (a) Harris, (b) KLT, (c) SIFT.*

Détection d'objet basée sur la segmentation de l'image

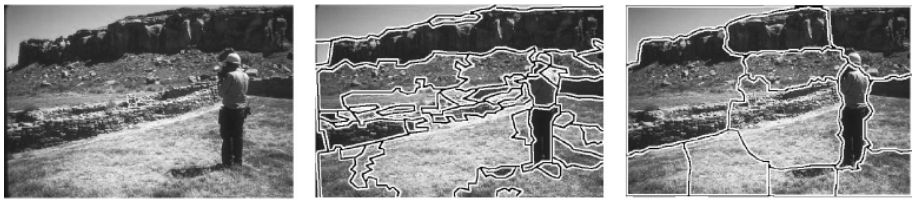
Le but des algorithmes de segmentation d'image est de partitionner l'image en régions perceptuellement similaires (figure 4.3). Les techniques récentes de détections d'objets basées sur la segmentation d'image utilisées pour le suivi d'objets sont les suivantes : segmentation basée sur la méthode du Mean-Shift de Comaniciu et Meer [25], segmentation utilisant les *Graph-Cuts* comme Shi *et al.* [92] et segmentation fondée sur les contours actifs.

La segmentation d'image utilisant l'approche du Mean-Shift se base sur l'estimation d'une densité à noyau dont la taille doit être définie préalablement. C'est une méthode itérative de montée de gradient convergeant vers le mode, c'est-à-dire le maximum local, de la densité d'un nuage de points. Comaniciu et Meer se placent à la fois dans le domaine spatial et colorimétrique. Cette méthode a l'avantage de ne reposer sur aucun a priori sur la distribution des intensités des pixels. Par contre, elle a tendance à générer un grand nombre de régions (figure 4.3(b)).

La méthode de segmentation par *Graph-Cuts* consiste à agréger les points les plus « proches » entre eux en se basant, par exemple, sur des primitives couleurs ou de mouvement. Cette approche descendante établit pour cela une matrice d'adjacence décrivant la similarité (poids) entre chacun des pixels (noeuds). On partitionne ensuite le graphe à l'aide d'un certain critère. La figure 4.3(c) montre la segmentation obtenue avec ce type de méthode.

Dans un contexte de contour actif, on suppose connu le type de l'objet recherché et donc la représentation de son contour. La segmentation est achevée lorsque un contour fermé devant représenter la frontière de l'objet englobe au plus proche la région définie par l'objet.

La détection d'objets par segmentation d'image fait partie des approches statique. L'utilisation de ce type de méthodes pour du suivi d'objets est possible en rajoutant la dimension temporelle en utilisant des contraintes issues des résultats de détection des images précédentes.



(a) (b) (c)

Figure 4.3: *Illustration de détection de zone par segmentation de l'image (a) [111] : (b) segmentation basée sur la méthode du Mean-Shift, (c) segmentation utilisant les Graph-Cuts.*

Détection d'objet basée sur une modélisation du fond

La détection d'objet peut se faire en construisant une représentation de la scène, classiquement appelée *modèle de fond*, puis en cherchant les variations entre le modèle et les nouvelles images de la scène. Toute détection de variation par rapport au modèle de fond représente un objet en mouvement. La difficulté de cette tâche est très variable selon les conditions d'acquisition et du contenu de la scène notamment due aux variations d'illumination : mouvement des nuages, de la végétation, les ombres, etc. Pour de plus amples détails sur les différentes méthodes existantes et notamment sur leurs performances, on peut se reporter aux études proposées par Piccardi [82] et Porikli [83].

Les méthodes de détection de mouvement basées sur la différence d'image entre des images consécutives sont étudiées depuis les années soixante-dix [51]. La méthode consiste à considérer, qu'à chaque instant t, l'image à l'instant $t-1$ représente l'arrière-plan de la scène et que les zones en mouvement sont celles qui ont changé d'apparence entre $t-1$ et t. La détection de mouvement est donc obtenue par dérivation temporelle d'une séquence d'images. La dérivée est très rapide à calculer mais elle est également très instable du fait de sa sensibilité à tout type de bruit. De plus, seul le passé immédiat est pris en compte : les mouvements lents ou saccadés sont mal détectés. Pour illustrer cette méthode, on peut citer les travaux de Lipton *et al.* [61]. Après avoir effectué la différence entre deux images consécutives, ils appliquent un seuil afin de déterminer les zones où il y a eu des changements dans la scène. Après une analyse en composantes connexes, ils classifient les zones extraites en régions de mouvement ou non.

Les méthodes dites de soustractions de fond consistent à construire un modèle de fond de la scène puis à appliquer une fonction de décision sur chaque nouvelle trame afin de décider, pour chaque pixel, si celui-ci appartient à l'arrière-plan de la scène ou à un objet mobile. Ce type de méthodes est devenu très populaire avec les travaux de Wren *et al.* [108]. Afin d'apprendre graduellement les variations dans le temps de la scène, ils proposent de modéliser la couleur de chacun des pixels de l'arrière-plan par une gaussienne. Les paramètres de moyenne et de variance sont appris à partir des observations couleur effectuées sur plusieurs images consécutives. Ensuite, pour toute nouvelle image, la ressemblance couleur entre le modèle et les pixels de l'image est calculée. Si un pixel est trop différent, il est marqué comme ne faisant pas parti du fond. Cependant, la modélisation par une unique gaussienne

est insuffisante pour modéliser des variations répétitives de couleur dues, par exemple, à l'ombre ou au reflet. Pour pallier à ce problème, Stauffer et Grimson [96] proposent de modéliser chaque pixel par un mélange de gaussiennes. Dans cette méthode, pour chaque pixel de l'image courante, on compare la couleur du pixel à chaque gaussienne du modèle jusqu'à ce qu'une correspondance soit trouvée ou non. Si une correspondance est trouvée, la gaussienne est mise à jour sinon une nouvelle gaussienne est créée avec la moyenne du pixel observé et une valeur initiale pour la variance.

Outre l'utilisation de l'information couleur, une autre approche est d'incorporer, en plus, une information spatiale de la scène par région. Elgammal *et al.* [31] proposent une méthode de soustraction de fond non-paramétrique utilisant l'estimation de densité des fenêtres de Parzen pour représenter le fond. Dans l'image courante, un pixel est comparé par rapport au modèle de fond mais aussi par rapport à ses plus proches voisins. Avec l'algorithme « *WallFlower* », Toyama *et al.* [102] proposent de traiter la soustraction de fond à trois niveaux : pixel, région et image. Chen *et al.* [23] proposent de traiter l'image par blocs de pixels afin d'améliorer le temps de calcul de la méthode. Ce type d'approche a l'avantage d'être moins sensible aux mouvements locaux et plus à même de traiter les cas où l'arrière plan est non-stationnaire.

Une autre approche alternative est de représenter les variations d'intensité d'un pixel dans une séquence d'images comme étant un état discret correspondant à des évènements dans la scène. Par exemple, lors d'un suivi de voiture, un pixel peut appartenir soit à l'arrière-plan, soit à une zone en mouvement, soit à de l'ombre. Rittscher *et al.* [89] utilisent les modèles de Markov cachés pour classifier de petits blocs de l'image en l'un des trois états cités précédemment. Stenger *et al.* [98] utilisent aussi les modèles de Markov cachés pour détecter si une lumière est allumée ou non dans une pièce.

Plus récemment, Porikli et Tuzel [85] proposent de modéliser le fond par un apprentissage bayésien récursif en estimant non seulement la moyenne et la variance, mais aussi la distribution de probabilité de la moyenne et de la covariance de chaque modèle. Cet apprentissage préserve la multimodalité du fond et, en outre, il est capable d'estimer le nombre de niveaux nécessaires pour représenter chaque pixel.

La plupart des méthodes de soustraction de fond ont été développées dans le cas de caméras statiques. Ces méthodes peuvent être appliquées à des vidéos acquises par des caméras dynamiques pour de petits déplacements entre les images successives et dans l'hypothèse où la scène est plane. Dans le cas de caméras ayant un mouvement de rotation pure, la modélisation de la scène peut se faire en créant un panoramique de la scène en acquérant des images successives durant la rotation de la caméra [94]. Les matrices d'homographie peuvent être utilisées pour décrire les transformations entre les différentes images.

Détection d'objet basée sur une reconnaissance de forme

Les systèmes de détection d'objet peuvent aussi s'appuyer sur des techniques de reconnaissance de forme. Le principal intérêt de ces techniques est qu'elles sont efficientes avec

des caméras mobiles, des objets fixes. De plus, elles permettent de séparer les objets d'intérêt et différencier différentes classes d'objet (personne, véhicule, bagage, etc). Ces méthodes de détection reposent sur l'apprentissage automatique de différentes vues de l'objet à partir d'exemples positifs (images d'une personne) et négatifs (images ne représentant pas de personne). Cela revient à classifier les détections selon un certain nombre de labels : par exemple, un label personne ou autre objet, personne assise, debout, couchée, etc.

La sélection des primitives joue ici un rôle important dans les performances de la classification car c'est elles qui vont permettre de distinguer une classe d'objet d'une autre. En plus des primitives évoquées précédemment, on trouve aussi, comme type de primitives utilisées, la surface d'un objet, son orientation, son apparence sous la forme d'une fonction de densité comme un histogramme. Une fois que le choix des primitives est effectué, les diverses apparences de l'objet sont apprises par une méthode d'apprentissage supervisée. Ces approches incluent les réseaux de neurones [90], l'*adaptative boosting* de Viola *et al.* [106], les arbres de décision [38] et les *Support Vector Machine* [78].

4.1.3 Méthodes de suivi d'objets

Le but du suivi d'objet est de générer la trajectoire de l'objet à tout instant en localisant sa position pour chaque image de la vidéo. Les tâches de détection et d'établissement de correspondance image à image peuvent se faire séparément ou en même temps. Dans le premier cas, pour chaque image de la vidéo, la position moyenne des objets est estimée par une méthode de détection puis la méthode de suivi permet de faire l'association temporelle. Dans le second cas, le positionnement et la correspondance d'une image à l'autre sont estimés conjointement par une mise à jour itérative de la position et de l'information du voisinage de l'objet à chaque image.

Deux types d'approches pour associer temporellement les détections sont utilisés : une approche déterministe ou une approche probabiliste. Nous allons développer ces différentes possibilités.

Approche déterministe

L'espace de recherche le plus simple est l'image entière (recherche exhaustive). Une telle recherche est très coûteuse en temps de calcul. De plus, aucune cohérence au niveau des cibles n'est garantie et l'algorithme de suivi peut sans distinction s'accrocher sur n'importe quel maximum local du critère de similarité.

Dans certains cas, la recherche exhaustive peut cependant être implémentée de manière très efficace grâce aux images intégrales [105] [84]. Cette technique est adaptée aux situations où l'évaluation du critère dans une fenêtre rectangulaire implique une somme sur tous les pixels de la région d'intérêt. Ainsi, en pré-calculant cette somme pour tous les pixels de l'image (intégrale du premier pixel de l'image au pixel courant), on peut calculer la valeur du critère dans une fenêtre donnée en quatre opérations arithmétiques. [86] propose une

méthode de suivi par recherche exhaustive, basée sur les matrices de covariances stockées dans des images intégrale approchant le temps réel.

Afin d'accélérer le traitement, la recherche peut se faire dans un voisinage de la position de la cible à l'instant précédent. La recherche d'association entre deux images consécutives peut être vue comme un problème d'optimisation combinatoire répondant à une certaine combinaison de contraintes. Ces contraintes sont liées notamment aux points suivants :

- *Proximité.* On suppose que la position de l'objet ne change pas notablement d'une image à l'autre.
- *Vitesse maximale.* Elle limite la zone de recherche d'association à un voisinage circulaire autour de l'objet.
- *Petit changement de vitesse.* Cela suppose que la direction et la vitesse de l'objet ne peuvent pas changer radicalement.
- *Mouvement similaire.* On impose que, dans un petit voisinage de l'objet, la vitesse soit similaire à celle de l'objet. Cette contrainte convient particulièrement dans le cas d'objet caractérisé par une multitude de points.

Une autre approche consiste à guider la recherche selon l'optimisation d'une fonction de coût. La position de la cible est alors déplacée vers un optimum de cette fonction. Un exemple très populaire de ce type d'approche est le suivi basé sur le Mean-Shift [25]. L'algorithme de suivi proposé par Comaniciu *et al.* [26] consiste à optimiser le coefficient de Bhattacharya calculé entre l'histogramme couleur de référence et l'histogramme de la cible candidate à l'instant t en utilisant un noyau pour pondérer les données. Cette méthode est relativement efficace : le maximum est atteint le plus souvent en quelques itérations seulement. Cependant, le principal problème de cette méthode est le risque de tomber dans des maxima locaux.

Approche probabiliste

Les mesures obtenues à partir des capteurs contiennent inévitablement du bruit. Les méthodes basées sur des associations statistiques résolvent le problème de suivi en prenant en compte les mesures et le modèle d'erreur durant l'estimation de l'état de l'objet.

Le problème de suivi d'un objet dans une image peut être formulé de façon probabiliste à l'aide de modèles dynamiques. Ces méthodes se composent classiquement de deux processus : un processus d'état X à valeurs dans $\mathbb{R}^n$, modélisant les propriétés de l'objet telles que la positon, la vitesse et l'accélération et un processus de mesure (ou encore d'observation) Z à valeurs dans $\mathbb{R}^p$.

On note x_k le vecteur d'état à l'instant k. L'équation d'évolution de x_k s'écrit de manière générale de la façon suivante :

$$x_k = f_k\left(x_{k-1}, v_{k-1}\right) \tag{4.1}$$

où f_k est une fonction éventuellement non-linéaire et v_{k-1} est un bruit blanc indépendant des états passés et présents.

Le vecteur d'état x_k est estimé à partir d'observations z_k régi par la loi suivante :

$$z_k = h_k\left(x_k, w_k\right) \tag{4.2}$$

où h_k représente une fonction d'observation éventuellement non-linéaire et w_k est un bruit blanc indépendant des états passés et présents. Les observations z_k sont indépendantes entre elles conditionnellement à l'état caché x_k.

On peut considérer ce modèle comme un modèle de Markov caché où les états cachés $\{x_k\}$ forment une chaîne de Markov et les observations $\{z_k\}$ vérifient l'hypothèse de canal sans mémoire, illustrée par la figure 4.4. Le problème de filtrage consiste alors à estimer le vecteur aléatoire x_k à l'instant k connaissant une suite d'observations $z_{1:k} = (z_1, ..., z_k)$. D'un point de vue formel, il s'agit d'estimer la densité de probabilité a posteriori $p(x_k|z_{1:k})$.

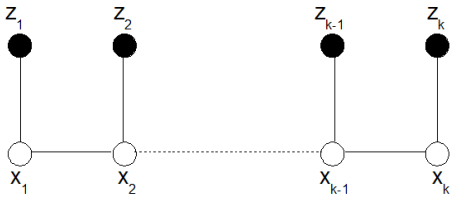


Figure 4.4: *Graphe de dépendance de la chaîne de Markov caché constituée par les processus x_k et z_k.*

Une solution théorique optimale est fournie par un filtre Bayésien récursif qui détermine $p(x_k|z_{1:k})$ en deux étapes : une étape de prédiction et une étape de mise à jour [1]. L'équation de prédiction relie la densité de probabilité a priori $p(x_k|z_{1:k-1})$ et la densité de probabilité a posteriori $p(x_{k-1}|z_{1:k-1})$ à l'instant précédent :

$$p(x_k|z_{1:k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|z_{1:k-1})dx_{k-1} \tag{4.3}$$

A l'instant courant, une nouvelle mesure z_k est disponible. Cette mesure permet, durant l'étape de mise à jour, de corriger la densité de probabilité prédite $p(x_k|z_{1:k-1})$ selon la règle de Bayes afin d'obtenir la densité de probabilité a posteriori $p(x_k|z_{1:k})$:

$$p(x_k|z_{1:k}) = \frac{p(z_k|x_k)p(x_k|z_{1:k-1})}{p(z_k|z_{1:k-1})} \tag{4.4}$$

Dans le cas d'un suivi mono-cible, si les fonctions f_k et h_k sont linéaires et les bruits v_k et w_k gaussiens, la solution optimale est donnée par le filtre de Kalman [17]. Dans le cas général, c'est-à-dire lorsque les variables ne suivent pas une distribution gaussienne, les filtres à particules [99] permettent de passer outre cette limitation.

Si l'on suit plusieurs objets avec des méthodes de suivi mono-cible tel que le filtre de Kalman ou les filtres particulaires, le problème d'association des mesures avec un objet en particulier va se poser. La méthode la plus simple est d'utiliser l'approche du plus proche voisin. Cependant, si un objet est très proche d'un autre objet, la correspondance peut être

incorrecte ce qui peut empêcher le filtre de converger. Plusieurs techniques statistiques d'association existent pour résoudre ce problème. Une étude détaillée de ces techniques peut être trouvée dans le livre de Bar-Shalom et Fortmann [2] ou dans l'étude menée par Cox [28]. Le filtre probabiliste à association de données (JPDAF en anglais) et le filtre à hypothèses multiples (MHT en anglais) sont les techniques usuellement utilisées.

4.2 Suivi d'objet multi-caméras

Le suivi d'objet dans un réseau de caméras est une extension du cas mono-caméra. Il contient intrinsèquement un suivi d'objet dans chaque caméra où l'on retrouve les techniques décrites dans la partie précédente.

L'utilisation d'un réseau de caméras répond à deux besoins. Tout d'abord, l'augmentation du nombre de caméras permet l'utilisation de l'information de profondeur grâce à la triangulation des données afin de résoudre des problèmes d'occultation ([33], [56]). Ensuite, à la différence des systèmes mono-caméra limités par un champ de vue fini, les systèmes de surveillance composés de plusieurs caméras peuvent contrôler de vastes zones. Dans ce cas, la plus grande difficulté est d'établir les correspondances pour une cible donnée entre les différentes vues et d'intégrer les informations venant des différentes caméras pour obtenir un suivi robuste face aux occultations sur l'ensemble de la scène observée.

Dans les paragraphes qui vont suivre, on va aborder plus en détail la complexité du passage de l'objet d'une caméra à une autre dans le cas où les caméras ont un champ de vue joint et dans le cas plus difficile où les champs de vue sont disjoints. Ensuite, on s'intéressera aux méthodes d'intégration des informations venant des algorithmes de suivi de chaque caméra : on parlera dans ce cas de fusion de données.

4.2.1 Suivi d'un objet du champ de vue d'une caméra à l'autre

Champ joint

Dans cette catégorie, une première classe d'approche suppose que l'objet est vu en permanence par plusieurs caméras calibrées. Il est alors possible de trianguler les différents points de vue et donc de réaliser un suivi 3D. Ainsi, Chang et Gong [21] combinent contraintes géométriques et modèles d'apparence pour gérer le suivi multi-cibles dans un système à deux caméras. Dans [11], les objets sont détectés par une soustraction de fond dans chaque image et mis en correspondance grâce aux contraintes géométriques (épipolaire). Le filtre de Kalman est rendu linéaire par l'optimisation des paramètres 3D.

Une seconde catégorie de système concerne les réseaux de caméras ayant pour but de couvrir une large superficie avec un minimum de caméras. Dès lors le champ commun entre les caméras se réduit aux transitions en bordure de champ. Le challenge est alors de gérer les transitions entre les champs de vue des caméras. Pour intégrer la connaissance de ces transitions, une étape de calibrage est nécessaire. Ce calibrage peut être fait hors-ligne à l'aide de mire ou par apprentissage en début de processus en analysant les trajectoires des

objets présents dans le champ des caméras. Ainsi, dans [55], les auteurs ne réalisent pas de calibrage fort de leurs caméras. Seules les lignes apprises à partir de quelques mises en correspondance délimitent les zones communes entre caméras. Dans [75], Nam *et al.* décrivent le réseau de caméras par un graphe non orienté dont les arcs sont les probabilités de transitions inter-caméras. Ces probabilités sont estimées à partir de résultats de suivi. Ce type d'approche suppose que dans la phase d'apprentissage toutes les transitions possibles soient effectuées une fois, ce qui reste hypothétique dans les réseaux complexes.

Dans [19], Cai et Aggarwal effectuent le suivi d'objet dans une seule vue tant que le sujet n'est vu que par une seule caméra. Lorsque le système détecte la cible dans une seconde caméra, le système exploite les différentes vues de la cible à travers une mise en correspondance exploitant la géométrie épipolaire entre images. Le système rebascule automatiquement en suivi mono-caméra lorsque l'objet occupe un champ suffisant dans une caméra afin de réduire le coût de calcul.

Champ disjoint

Dans le cas où les champs entre les différentes caméras sont disjoints, la problématique de suivi devient quasiment un problème pur de reconnaissance d'objet du fait du relâchement des contraintes spatiales. En effet, n'ayant plus de continuité d'une image à l'autre, il faut arriver à retrouver exactement le même objet dans une nouvelle configuration de la scène par rapport à l'instant précédent.

La plupart des travaux sur ce sujet consistent à créer des dépendances entre caméras pour passer d'une caméra à une autre. Des solutions sont proposées pour calibrer automatiquement un réseau de caméras à champ disjoint à partir d'une mesure statistique des dépendances entre caméras ([36], [100]). Deux caméras sont considérées comme connectées si un objet sortant du champ de vue d'une caméra est vu entrant dans l'autre. Cette dépendance est caractérisée par la distribution de transformation d'observations telle que le temps de passage ou la couleur. Madden *et al.* [65] modélisent chaque cible par un descripteur couleur (basé sur les K-Means) permettant de réaliser les mises en correspondance entre caméras. Ces approches restent très dépendantes des dimensions des zones non couvertes. La combinatoire peut devenir difficile à gérer si le nombre d'objets augmente ou si la topologie du réseau devient complexe. Dans [74], une solution est proposée pour suivre des véhicules dans un réseau de caméras non joint. L'association des données est réalisée à partir d'un critère de raisonnement temporel et d'un critère d'apparence visuelle de chaque objet (dimension et couleur). Les contraintes temporelles intègrent un modèle d'accélération des véhicules mis à jour en fonction des conditions du trafic.

4.2.2 Fusion de données hétérogènes

Selon l'étude de Luo *et al.* [64], les méthodes de fusion de données hétérogènes sont classifiées comme suit :

◇ *Méthodes basées sur l'estimation*

La moyenne pondérée des informations redondantes provenant du groupe de capteurs est une méthode simple et intuitive de fusion de données. Bien que cette méthode permette de traiter en temps-réel des données bas-niveau, le filtre de Kalman est largement préféré car les résultats d'estimation de données fusionnées sont optimaux au sens statistique dans le cas d'une modélisation linéaire du système et gaussienne de l'erreur. Dans le cas d'une modélisation non-linéaire, les filtres à particules sont utilisés.

◇ *Méthodes de classification*

Les méthodes de classification en cluster sont un puissant outil, qui s'appuient trois types d'approches. Les méthodes hiérarchiques d'agglomération s'appuient sur la construction d'une classe pour chaque mesure et fusionnent les classes selon une mesure de distance jusqu'à ce que soit atteint un nombre prédéfini de classes ou que la distance soit trop importante entre deux classes. Les méthodes hiérarchiques de division commencent avec une seule classe qui est ensuite divisée en deux ou plusieurs classes selon une mesure de distance jusqu'à ce que soit atteint un nombre prédéfini de classes ou que la distance maximale soit atteinte entre deux classes. Les méthodes de partitionnement itératives fonctionnent sur un nombre fixe de classes. Les mesures sont distribuées de manière aléatoire sur le groupe de classes. Celles-ci sont ajustées jusqu'à ce que la moyenne de chaque classe à l'intérieur de la distance soit minimale.

◇ *Méthodes d'inférence*

Les méthodes d'inférence bayésienne permettent de représenter l'information sous la forme de densité de probabilité selon les règles de la théorie probabiliste. On obtient ainsi une relation entre la probabilité a priori d'une hypothèse, la probabilité conditionnelle d'une observation donnant une hypothèse et la probabilité a posteriori de l'hypothèse. Ces modèles sont néanmoins assez limités lorsque des opinions conflictuelles doivent être fusionnées ou lorsque l'information a priori est inconnue dans le cas de la théorie de Bayes.

Les modèles basés sur la théorie de l'évidence [30] sont adaptés pour la prise en compte de l'incertain et de l'imprécision. Dans la théorie de Dempster-Shafer, à partir d'une structure de croyance, il est possible de définir la crédibilité et la plausibilité de chaque proposition, qui sont l'expression respective de la borne inférieure et supérieure de la probabilité de cette proposition. Une structure de croyance correspond donc à une distribution d'intervalles de probabilités. La crédibilité mesure à quel point les informations données par une source soutiennent la proposition. La plausibilité mesure à quel point les informations de données par une source ne contredisent pas la proposition.

◇ *Méthodes d'intelligence artificielle*

Un réseau de neurones est un modèle de calcul dont la conception est très schématiquement inspirée du fonctionnement de vrais neurones. Les réseaux de neurones, en tant que système capable d'apprendre, mettent en œuvre le principe de l'induction, c'est-à-dire l'apprentissage par l'expérience. Par confrontation avec des situations ponctuelles, ils infèrent un système de décision intégré dont le caractère générique est fonction du nombre de cas d'apprentissages rencontrés et de leur complexité par rapport à la complexité du problème à résoudre.

La logique floue (*fuzzy logic*, en anglais) est un ensemble de théories mathématiques qui traite de la représentation et de la manipulation de connaissances imparfaites (imprécises, incertaines voire incomplètes). Dans la théorie classique des ensembles, un élément d'un ensemble de référence peut, soit ne pas appartenir (0), soit appartenir (1) à un sous-ensemble. L'appartenance est donc pensée de manière binaire. La théorie des sous-ensembles flous assouplit cette notion d'appartenance en la relativisant, en la définissant par un nombre réel (appelé niveau d'appartenance) allant de 0 (pas du tout) à 1 (absolument).

Dans la littérature concernant le suivi de personne, les méthodes basées sur l'estimation sont les plus utilisées pour fusionner des données venant de sources différentes (algorithme et/ou capteur). En effet, en pratique, plusieurs facteurs contribuent à amener de l'incertitude sur la position et la configuration exactes de l'objet à suivre. Ceci inclut les bruits des observations, une modélisation inexacte, les faux positifs, les fluctuations dans les conditions environnementales. Ce type de méthode permet de suivre des objets divers de façon robuste malgré des changements majeurs et rapide de pose et d'illumination. Afin d'illustrer ces propos, on va présenter quelques travaux de suivi basé sur des méthodes probabilistes soit pour faire du suivi mono-capteur multi-critères soit pour faire du suivi multi-capteurs multi-critères.

Suivi mono-capteur multi-critères

Le filtrage particulaire consiste en une approximation séquentielle de type Monte Carlo à base d'échantillonnage pondéré d'une part (étape de « propagation » des échantillons précédents selon une fonction d'importance et pondération), et de ré-échantillonnage d'autre part. Si tous les ingrédients sont anciens, leur combinaison effective pour résoudre les problèmes de filtrage non linéaire est récente ([48], [57]). Dans les approches particulaires, l'échantillonnage, dit « d'importance », est une manière de diriger la recherche en combinant une information de prédiction fondée sur la position précédente de l'objet et sur le mouvement avec une connaissance supplémentaire qui peut être disponible à partir des capteurs. L'échantillonnage d'importance s'applique lorsqu'une connaissance supplémentaire est disponible sous la forme d'une fonction d'importance qui décrit les parties de l'espace d'état qui contiennent le plus d'informations sur la distribution de probabilité a posteriori. L'idée est de concentrer les particules dans ces zones de l'espace d'état en les générant à partir de la fonction d'importance. Cette technique évite de générer des échantillons de faible poids qui sont éliminés dans l'étape de ré-échantillonnage. Plusieurs travaux utilisant cette caractéristique des filtres à particules vont être présentés.

Dans [49], l'application concerne le suivi de main. Isard et Blake disposent d'images d'apprentissage de mains à partir desquelles ils expriment la probabilité pour un pixel donné d'être de la couleur de la peau. Ils effectuent alors une segmentation au sens de la couleur en repérant les régions de la couleur de la peau. Cette connaissance est utilisée pour construire la fonction d'importance qui guide, ensuite, la recherche de la cible « mains »vers les régions de couleur de la peau. Leur densité de probabilité s'écrit comme un mélange

pondéré de gaussiennes. Ainsi, les auteurs obtiennent un suivi efficace même lorsque la main effectue un mouvement rapide. En effet, en cas de mouvement soudain qui n'est pas prédit par le modèle dynamique, la segmentation détecte la nouvelle position de la main et des particules du filtre particulaire sont générées aux alentours de cette position permettant le suivi de mouvement. L'introduction de la segmentation couleur dans l'échantillonnage d'importance offre, par ailleurs, la possibilité de réinitialiser le suivi en cas de perte de la cible. Ainsi, en utilisant des sources complémentaires, les auteurs parviennent à un suivi robuste.

Afin de suivre plusieurs joueurs de hockey, Okuma *et al.* [76] utilisent une fonction d'importance fondée sur un module de détection des joueurs de hockey. La distribution de proposition est construite à parti d'un mélange qui intègre, d'une part, l'information des modèles dynamiques de chaque joueur, et, d'autre part, les hypothèses de détection des joueurs générés par un module de type *Adaboost*. Ce mélange permet non seulement de conserver la trace de chaque joueur grâce au modèle dynamique mais aussi de détecter rapidement les joueurs entrant dans la scène au moyen d'Adaboost [106] et d'initialiser automatiquement le suivi pour ces nouveaux objets.

Dans [18], Bullock et Zelek proposent d'utiliser la complémentarité d'informations de mouvement et de couleur. Comme les algorithmes perdent leur cible pendant les périodes de mouvement et d'occultations, ils intègrent l'information de mouvement dans la fonction d'échantillonnage d'importance pour retrouver la cible qui correspond effectivement à une zone de mouvement. L'approche de suivi se focalise finalement sur la bonne région grâce à l'information de couleur.

Dans [8], Bichot *et al.* proposent, dans la cadre d'une application de reconnaissance de poissons , d'exploiter le résultat d'une segmentation par étude du mouvement afin d'améliorer le suivi par filtrage particulaire. Les hypothèses du filtre à particules correspondent aux régions dont le mouvement est similaire au déplacement de la cible à l'instant précédent. Ensuite, après avoir identifié la région de mouvement correspondant à la cible, ils apprennent le modèle de la cible qui est intégré comme référence lors de la prochaine itération du suivi. Ils montrent qu'ainsi les performances du suivi d'objets non-rigides et animés de mouvements complexes dans un milieu perturbé sont améliorées par rapport à l'utilisation d'un filtre Boostrap conventionnel.

Suivi multi-capteurs multi-critères

Dans [54], Kang *et al.* proposent une nouvelle approche pour intégrer des informations venant à la fois de caméras statiques et de caméras dynamiques. Celle-ci consiste à traiter les trajectoires observées par chaque caméra avec un filtre de Kalman. Le suivi de personnes est basé sur deux modélisations : une modélisation de l'apparence et une modélisation du mouvement. Le résultat final du suivi, à chaque instant, est le maximum de la probabilité jointe basé, entre autre, sur le résultat du filtre de Kalman appliqué à l'information d'apparence d'une part, et à l'information de mouvement d'autre part :

$$P_{total}(X_t) = P_{apparence}(X_t) P_{mouvement}(X_t) P_{total}(X_{t-1}). \tag{4.5}$$

Cette approche leur permet d'obtenir un suivi plus précis et plus robuste aux occultations et détections incomplètes.

Dans [109], Yao *et al.* utilisent un système de vision basé sur une caméra omnidirectionnelle et une caméra Pan-Tilt-Zoom. Afin de faire un suivi basé sur la coopération des deux caméras, ils proposent de fusionner les données venant de chaque caméra grâce à des filtres de Kalman distribués. L'objectif de la coopération est donc d'obtenir une estimation du vecteur d'état à un instant donné en maximisant la probabilité a posteriori basée sur les séquences d'observations de chaque caméra. Le filtre de Kalman est divisé en deux étapes : une étape de prédiction et une étape de mise à jour. Yao *et al.* ont testé l'échange d'information dans chacune des étapes. Ils concluent à la fin de leur étude que le meilleur résultat est obtenu lorsque l'échange d'information se fait lors de l'étape de prédiction. Grâce à cette coopération entre les deux capteurs, les auteurs parviennent à un suivi précis.

Han *et al.* [39] proposent une méthodologie de modélisation des interactions entre les différents capteurs basée sur une mixture de filtres Bayesiens, un filtre par capteur. Afin de garder les caractéristiques multi-modales de chaque capteur, au lieu de combiner les données de chaque capteur lors de l'étape de mesure, ils proposent de fusionner les données lors de l'étape de mise à jour. La probabilité a posteriori combinée est construite sur une mixture de probabilité a posteriori individuelle. La probabilité a posteriori d'un capteur contribue à la probabilité a posteriori combinée en fonction de la confiance dans l'observation issue de chaque capteur. Han *et al.* montrent dans leur étude qu'ainsi l'estimation de la probabilité a posteriori est plus précise.

Dans [88], Pérez *et al.* considèrent deux types de mesures afin de suivre le visage du locuteur : d'une part, la couleur dans l'image, et, d'autre part, les pics de corrélation entre les signaux sonores enregistrés par deux microphones disposés de part et d'autre de la caméra. Le but est de suivre dans l'image le visage du locuteur. Dans ce cas, l'approche particulaire ne nécessite pas d'opération de triangulation explicite. Il suffit d'évaluer qu'elle est, au regard des signaux sonores reçus, la vraisemblance de l'hypothèse de localisation associée à chaque particule. Le bénéfice de la fusion résulte de la complémentarité des modalités : une modalité relative à l'apparence de l'objet suivi (couleur), persistante mais sujette à confusion, est combinée à une seconde (le son), intermittente mais précise lorsqu'elle est présente.

Chapitre 5

Contributions au suivi haute résolution d'une personne à l'aide d'un système de vision hybride

A l'aide du système de vision hybride composé d'une caméra statique et d'une caméra dynamique, on souhaite suivre une personne avec la meilleure résolution possible, c'est-à-dire en adaptant le zoom de la caméra dynamique à la taille de la cible suivie.

Dans un premier temps, nous avons choisi de valider notre méthode de calibrage pour une application de suivi en mettant en place un système maître-esclave : le suivi de la cible s'effectue dans la caméra statique qui commande à chaque itération la caméra dynamique afin d'obtenir un suivi haute résolution de la personne.

Dans un second temps, un système de suivi basé sur une collaboration plus étroite entre les deux capteurs est proposé afin de résoudre les limites d'un système maître-esclave.

Ces différents points seront abordés dans ce chapitre. Les résultats de chaque méthode seront présentés dans le chapitre 6.

5.1 Suivi mono-cible par un système maître-esclave

Deux types de système maître-esclave basés sur la combinaison d'une caméra statique et d'une caméra dynamique existent dans la littérature. Ils sont illustrés par les travaux récents de Zhou *et al.* et de Bodor *et al.* .

Dans [112], Zhou *et al.* proposent un système de vidéo-surveillance combinant une caméra Pan-Tilt-Zoom et une caméra statique. La caméra statique sert à initialiser le suivi de personne effectué dans la caméra dynamique, approche illustrée par la figure 5.1. Ils commencent par une étape de détection afin de raffiner la position de la caméra dynamique pour qu'elle soit exactement centrée sur la personne. Ensuite, ils utilisent pour l'étape de suivi l'algorithme du Mean-Shift basé sur un modèle d'apparence d'histogramme couleur.

Afin d'avoir la puissance de calcul maximale, les auteurs ont réparti les divers traitements sur trois unités de calcul : une pour les traitements effectués dans la caméra statique, une pour les traitements effectués dans la caméra dynamique et une pour gérer les commandes envoyées à la caméra dynamique.

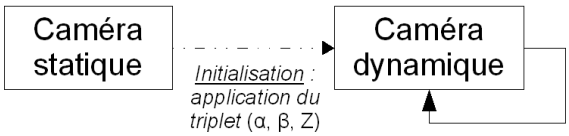


Figure 5.1: *Illustration du système maître-esclave utilisé par Zhou* et al. *[112].*

Bodor *et al.* [13] proposent d'utiliser, en continu, la caméra statique pour déterminer les paramètres de commande de la caméra Pan-Tilt-Zoom, illustrée par la figure 5.2. L'ensemble des traitements (détection et suivi) se fait au niveau de la caméra statique. La caméra dynamique est juste pilotée : son rôle se limite à prendre des images haute définition des cibles suivies.

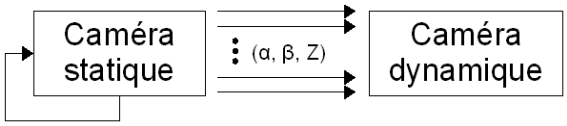


Figure 5.2: *Illustration du système maître-esclave utilisé par Bodor* et al. *[13].*

Par rapport au système de suivi proposé par Zhou *et al.* , pour le même type de résultat, le système de Bodor *et al.* nécessite moins de puissance de calcul et moins de traitement (seulement dans la caméra statique). De plus, cette méthode permet un système de contrôle simplifié puisque l'objet a une vitesse apparente plus lente dans l'image de la caméra statique et parce qu'une méthode de détection de cible en mouvement plus simple est effectuée dans la caméra statique.

L'approche proposée par Bodor *et al.* est basée sur un calibrage *fort* de la paire de caméras : calcul des paramètres intrinsèques et extrinsèques des deux caméras. C'est à partir de ces paramètres qu'ils déterminent la commande à appliquer à la caméra dynamique.

Par sa simplicité de mise en œuvre et afin de tester la méthode de calibrage proposée précédemment, on a choisi de mettre en place la même approche de système maître-esclave que Bodor *et al.* en utilisant notre solution de calibrage faible. Celle-ci donne une relation directe entre une position (x, y) dans l'image de la caméra statique et les paramètres angulaires à appliquer pour que la caméra dynamique soit centrée sur cette position. Il nous faut donc une méthode de détection et de suivi de cible qui permette de déterminer le centre de gravité de la cible pour pouvoir appliquer directement le résultat de notre calibrage et la taille de la cible afin d'estimer la commande de zoom.

Nous allons décrire dans la suite de cette partie les choix de méthodes utilisés dans notre application de suivi basée sur un système maître-esclave.

5.1.1 Détection d'objets en mouvement

On souhaite détecter l'ensemble des zones en mouvement de notre scène dans la caméra statique, puis amener la caméra dynamique vers une cible choisie (en adaptant le zoom à la taille de la cible). Pour que le système soit fonctionnel, la détection doit être à la fois précise et rapide, c'est pourquoi nous avons choisi une méthode de soustraction de fond, au moyen d'un mélange de gaussiennes.

En effet, Piccardi [82] nous assure que cette méthode donne des résultats très précis par rapport au modèle, tout en restant dans des temps de calculs raisonnables pour notre application (qui ne dépendent que du nombre de gaussiennes choisies).

La plupart des méthodes de soustraction de fond traitent les images pixel par pixel ([31], [85], [96]). Récemment, de nouvelles approches sont apparues traitant les images par blocs de pixels ([23], [69], [72]). Ces approches ont l'avantage d'être moins sensibles aux variations locales (figure 5.3) et sont plus à même de traiter les cas où l'arrière plan est non-stationnaire. De plus, l'approche par bloc permet une implémentation plus efficace, qui donne en pratique une détection en temps réel.

Figure 5.3: *Comparaison d'une méthode de soustraction de fond basée sur une approche pixel (image de gauche) et celle basée sur une approche par bloc (image de droite) comme Chen* et al. *[23].*

La méthode que l'on va détailler par la suite est basée sur les travaux de Chen *et al.* [23] qui proposent un descripteur discriminant appelé histogramme de contraste, extension des travaux de Huang *et al.* [46], pour décrire chaque bloc de pixels. Chaque bloc de l'image est décrit directement par un histogramme de contraste basé sur les couleurs des pixels du bloc. Puis, un mélange de Gaussiennes est utilisé pour mettre à jour le modèle. Les auteurs proposent une étape de raffinement au niveau des pixels mais celle-ci est coûteuse en temps de calcul.

Nous avons fait le choix de nous limiter à la modélisation par blocs. Le fait de ne pas faire de raffinement au niveau du pixel limite naturellement la détection aux objets de la taille du bloc de pixels (taille 8×8 pixels). Cette limitation n'étant pas contraignante dans notre cas, nous avons préféré privilégier le gain en terme de temps d'exécution car le

système dynamique de suivi doit respecter une contrainte temps réel forte.

Histogramme de contraste

◇ *Cas d'images en niveau de gris*
L'image est partitionnée en blocs de taille 8×8 pixels et un descripteur de contraste basé sur le calcul de gradient est construit pour chaque bloc. On note $\mathbf{p}$ un pixel du bloc $\mathbf{B}$ et $\mathbf{p_c}$ le pixel du centre du bloc $\mathbf{B}$. Comme $\mathbf{p_c}$ n'existe pas forcément, la valeur de $\mathbf{p_c}$ est estimée comme la moyenne des quatre pixels centraux de $\mathbf{B}$. Pour chaque pixel de chaque quadrant q_i de $\mathbf{B}$, les valeurs des histogrammes de contraste positives et négatives sont calculées.

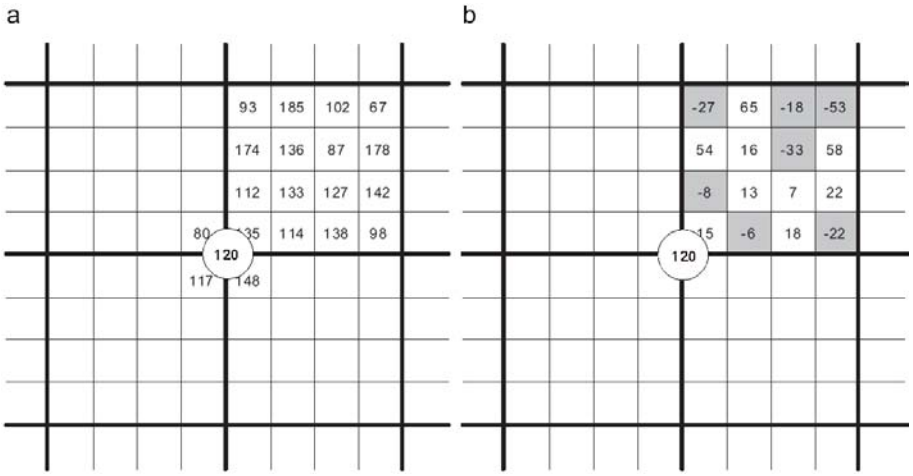


Figure 5.4: *Exemple de calcul, pour un pixel, des valeurs des histogrammes de contraste positives et négatives [23]. On suppose que le bloc est de taille 8×8 et que le point est le centre $\mathbf{p_c}$ du bloc. Chaque bloc est séparé en quatre quadrants. (a) Image originale en niveau de gris. La valeur moyenne de $\mathbf{p_c}$ est 120 ($= \frac{80+135+117+148}{4}$).(b) Les valeurs de contraste des pixels du premier quadrant ont été calculés : les valeurs négatives sont grisées, les valeurs positives étant sur fond blanc. Les valeurs des histogrammes de contraste positives, $CH^+_{q_0}(\mathbf{p_c})$, et négatives, $CH^-_{q_0}(\mathbf{p_c})$, sont respectivement : 29.79 ($= \frac{65+54+16+58+13+7+22+15+18}{9}$) et -23.86 ($= \frac{-27-18-53-33-8-6-22}{7}$)*

Pour le pixel du quadrant q_i, la valeur correspondante de l'histogramme de contraste positif $CH^+_{q_i}(\mathbf{p_c})$ du quadrant est définie par la relation suivante :

$$CH^+_{q_i}(\mathbf{p_c}) = \sum_{\mathbf{p_n^i} \in q_i \text{ et } n \in N^+} \frac{\mathbf{p_n^i} - \mathbf{p_c}}{N^+} \tag{5.1}$$

où N^+ est le nombre maximal de pixels de q_i tel que $\forall\ 0 \leq n \leq N^+, \left(\mathbf{p_n^i} - \mathbf{p_c}\right) \succ 0$. De même, on a pour le calcul des valeurs négatives de l'histogramme de contraste pour

le pixel du quadrant q_i la relation suivante :

$$CH^-_{q_i}(\mathbf{p_c}) = \sum_{\mathbf{p^i_n} \in q_i \text{ et } n \in N^-} \frac{\mathbf{p^i_n} - \mathbf{p_c}}{N^-} \tag{5.2}$$

où N^- est le nombre de pixel de q_i tel que $\left(\mathbf{p^i_n} - \mathbf{p_c}\right) \prec 0$.
Le descripteur de l'histogramme de contraste pour un bloc d'une image en niveau de gris est un vecteur de dimension 8 défini par la relation suivante :

$$CH_G(\mathbf{p_c})^T = \left(CH^+_{q_0}, CH^-_{q_0}, CH^+_{q_1}, CH^-_{q_1}, CH^+_{q_2}, CH^-_{q_2}, CH^+_{q_3}, CH^-_{q_3}\right). \tag{5.3}$$

Le calcul de ce descripteur est illustré par la figure 5.4.

◇ *Cas d'images couleur*
Dans le cas d'images couleur, on duplique la méthode pour les images en niveau de gris pour chaque canal de couleur. Soit $j \in \{R, G, B\}$ l'indice du canal de couleur pour le pixel $\mathbf{p}$ et $k \in \{R, G, B\}$ pour $\mathbf{p_c}$. Dans ce cas, les histogrammes de contraste de couleur positifs et négatifs sont respectivement déterminés par les relations suivantes :

$$CH^{j,k^+}_{q_i}(\mathbf{p_c}) = \sum_{\mathbf{p^i_n}^j \in q^j_i \text{ et } n \in N^+} \frac{\mathbf{p^i_n}^j - \mathbf{p_c}^k}{N^+} \tag{5.4}$$

et

$$CH^{j,k^-}_{q_i}(\mathbf{p_c}) = \sum_{\mathbf{p^i_n}^j \in q^j_i \text{ et } n \in N^-} \frac{\mathbf{p^i_n}^j - \mathbf{p_c}^k}{N^-} \tag{5.5}$$

où $\left(\mathbf{p^i_n}^j - \mathbf{p_c}^k\right)$ est la valeur de contraste entre le pixel $\mathbf{p}$ du canal de couleur j et $\mathbf{p_c}$ du canal de couleur k.
Il y a donc neuf combinaisons de paire (j, k) ce qui donne un vecteur de descripteur de taille 72 pour chaque bloc. Selon Chen *et al.* , six paires suffisent à construire un descripteur efficace, ce qui donne un descripteur de dimension 48.

La version utilisée est une extension des travaux de Chen *et al.* . Outre les paramètres utilisés par les auteurs, une description du pixel par sa couleur dans l'espace RGB a été ajoutée. On obtient un descripteur de dimension 51. Cette modification permet de différencier des zones uniformes de couleurs différentes, qui dans la version de base pouvaient être décrites par un même descripteur [81]. Par contre, la méthode perd un peu en robustesse face aux changements d'illumination.

Une fois la soustraction de fond effectuée, une labellisation rapide en composantes connexes (ou *blob* en anglais), basée sur la méthode proposée par Chang *et al.* [20], associée à des opérations de morphologie mathématique afin de filtrer le bruit, est appliquée afin d'extraire une liste de blobs.

Mélange de gaussiennes

Le mélange de gaussiennes, inspirée des travaux de Stauffer et Grimson [96], est construit à partir des descripteurs couleur de contraste de chaque bloc de l'image. A l'instant t, l'histogramme couleur de contraste de chaque bloc est modélisé par K distributions Gaussiennes avec un poids, noté $\omega_{k,t}$, tel que $\sum_{k=1}^{K} \omega_{k,t} = 1$. La probabilité pour l'instant courant $t+1$, que le bloc appartienne à l'arrière plan est

$$P(x_{t+1}) = \sum_{k=1}^{K} \omega_{k,t} \times \eta\left(x_{t+1}, \mu_{k,t}, \Sigma_{k,t}\right) \tag{5.6}$$

où $\mu_{k,t}$ et $\Sigma_{k,t}$ sont le vecteur moyenne et la matrice de covariance de la k^{ieme} gaussienne à l'instant t.

Chaque bloc est ensuite classé, sur la base de la distribution gaussienne, comme faisant partie de l'arrière plan ou comme objet en mouvement.

Résultats de la soustraction de fond

Figure 5.5: *Résultats de la soustraction de fond fondée sur la méthode d'histogramme couleur proposée par Chen* et al. *[23].*

Cette méthode de soustraction de fond nous permet donc de localiser les objets en mouvement dans la scène. Les informations extraites utilisables pour la commande de la caméra dynamique sont :

- ◇ la hauteur h_y et la largeur h_x de la boîte englobant l'objet détecté, en vert sur la figure 5.5,
- ◇ le centre de gravité de la boîte, marqué par une croix sur la figure 5.5.

On constate, dans le cas du couloir illustré par la figure 5.5, que les résultats de la détection par soustraction du fond sont satisfaisants même lorsque l'objet est relativement petit dans l'image.

5.1.2 Suivi de blob

La méthode de détection présentée précédemment donne, à chaque instant, une liste de blobs ordonnés selon leur taille. Il n'y a cependant aucune cohérence spatiale au cours du temps car la détection ne donne pas d'association de blobs entre deux images consécutives. Le problème est donc de savoir sur quel critère on va choisir le blob à suivre au cours du temps. La seule information que l'on a en sortie de la détection est une liste de positions et de tailles de blobs. Si l'on choisit comme critère de choix la taille du blob, par exemple celui qui a la plus grande surface, au cours du temps, vu la configuration de la scène, le blob qui a un instant t est considéré comme le plus gros ne le sera plus, à un instant $t' > t$, si la personne suivie se retrouve au fond du couloir et qu'une autre personne entre dans la scène au niveau des caméras. On s'aperçoit qu'il est nécessaire de mettre en place une méthode de suivi dans la caméra statique permettant de savoir d'une image à l'autre la nouvelle position du blob que l'on suit afin de piloter de manière cohérente dans le temps la caméra dynamique.

L'estimation de la trajectoire de la cible est effectuée dans l'espace de l'image donc en deux dimensions, bien que la modélisation des trajectoires soit difficile en 2D car un déplacement réel de la personne en 3D peut entraîner un déplacement virtuel important dans l'image. Nous avons choisi d'utiliser une méthode de suivi probabiliste qui permet d'obtenir une cohérence spatiale en affinant un peu les associations entre les observations et les trajectoires.

Notre choix s'est porté sur le filtre de Kalman étendu, parfois appelé filtre de Kalman étendu de Schmidt [53] qui est une extension du filtre de Kalman au cas des fonctions non linéaires. L'objectif est de prédire l'évolution d'un système et de corriger les prédictions en fonction des observations a posteriori.

Filtre de Kalman étendu

Le système à un instant k est caractérisé par un vecteur d'état qui regroupe un ensemble d'informations (coordonnées, vitesse, orientation, accélération, ...) sur le système. Comme il est impossible de connaître les valeurs réelles du vecteur d'état $\mathbf{x}_k$ et de l'observation $\mathbf{z}_k$ du système, on va s'intéresser à leurs valeurs estimées. Le filtre de Kalman étendu repose sur un développement au premier ordre des fonctions d'observations et d'évolution au voisinage des valeurs estimées $\hat{\mathbf{x}}_{k|k}$ et $\mathbf{x}_{k|k-1}$ afin de se ramener à un problème linéaire. Cependant, si l'amplitude de l'erreur entre la valeur estimée et la valeur réelle est importante, les approximations qui interviennent dans le filtre deviennent grossières et peuvent faire diverger le système. Malgré ce risque, nous avons choisi d'utiliser ce type de filtre, suffisant dans notre cas.

Par convention, le vecteur $\hat{\mathbf{x}}_{k|k}$ représente le vecteur d'état estimé à l'instant k en utilisant l'observation disponible à cet instant. Le vecteur $\hat{\mathbf{x}}_{k+1|k}$ est la prédiction de l'état faite à l'instant suivant $k+1$ à partir de l'état $\hat{\mathbf{x}}_{k|k}$. Le filtre de Kalman étendu est donné par les relations suivantes :

◦ Système :

Équation d'état $\mathbf{x}_{k|k-1} = f(\mathbf{x}_k, \mathbf{u}_k)$
Équation de mesure $\mathbf{z}_k = h(\mathbf{x}_k, \mathbf{v}_k)$

- Évolution de l'état :
  Vecteur d'état $\hat{\mathbf{x}}_{k|k-1} = f(\hat{\mathbf{x}}_k, \hat{\mathbf{u}}_k)$
  Matrice de covariance de l'état $\mathbf{P}_{k|k-1} = \mathbf{F}_k \mathbf{P}_{k|k} \mathbf{F}_k^T + \mathbf{Q}'_k$
- Mise à jour de l'état :
  Gain de Kalman $\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{H}_k^T \left[\mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}'_k\right]^{-1}$
  Vecteur d'état $\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} - \mathbf{K}_k \left(\mathbf{z}_k - h(\hat{\mathbf{x}}_{k|k-1}, \hat{\mathbf{v}}_k)\right)$
  Matrice de covariance de l'état $\mathbf{P}_{k|k} = \left[\mathbf{I} - \mathbf{K}_k \mathbf{H}_k\right] \mathbf{P}_{k|k-1}$

où $\mathbf{F}_k$, $\mathbf{Q}'_k$ et $\mathbf{Q}_k$ sont les matrices jacobiennes des dérivées partielles de f ainsi que les matrices $\mathbf{H}_k$, $\mathbf{R}'_k$ et $\mathbf{R}_k$ pour la fonction h telles que :

$$\mathbf{F}_k = \frac{\partial f(\mathbf{x}_k, \hat{\mathbf{u}}_k)}{\partial \mathbf{x}_k}_{\mathbf{x}_k = \hat{\mathbf{x}}_{k|k}} \qquad \mathbf{H}_k = \frac{\partial h(\mathbf{x}_k, \hat{\mathbf{v}}_k)}{\partial \mathbf{x}_k}_{\mathbf{x}_k = \hat{\mathbf{x}}_{k|k-1}}$$

$$\mathbf{Q}'_k = \frac{\partial f(\hat{\mathbf{x}}_{k|k}, \mathbf{u}_k)}{\partial \mathbf{u}_k}_{\mathbf{u}_k = \hat{\mathbf{u}}_k} \qquad \mathbf{R}'_k = \frac{\partial h(\hat{\mathbf{x}}_{k|k-1}, \mathbf{v}_k)}{\partial \mathbf{v}_k}_{\mathbf{v}_k = \hat{\mathbf{v}}_k}$$

$$\mathbf{Q}_k = \frac{\partial f(\hat{\mathbf{x}}_{k|k}, \mathbf{u}_k)}{\partial \mathbf{u}_k}^T_{\mathbf{u}_k = \hat{\mathbf{u}}_k} \qquad \mathbf{R}_k = \frac{\partial h(\hat{\mathbf{x}}_{k|k-1}, \mathbf{v}_k)}{\partial \mathbf{v}_k}^T_{\mathbf{v}_k = \hat{\mathbf{v}}_k}$$

Mise en pratique

En pratique, le modèle utilisé est défini par le vecteur d'état comportant quatre paramètres : la position (p_x, p_y) de la cible et la vitesse définie sur les deux axes (v_x, v_y). La vitesse est considérée constante entre deux instant k et $k + \Delta t$. On obtient la définition de la fonction f avec le système d'équation suivant :

$$\begin{cases} p_{x,k} = p_{x,k-1} + v_x \Delta t \\ p_{y,k} = p_{y,k-1} + v_y \Delta t \\ v_{x,k} = v_{x,k-1} \\ v_{y,k} = v_{y,k-1} \end{cases}$$

La matrice $\mathbf{F}_k$ s'écrit donc de la manière suivante :

$$\mathbf{F}_k = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.7}$$

Le modèle dynamique du filtre de Kalman étendu doit être suffisamment souple pour s'adapter à un système mal connu (du fait que l'on observe des déplacement en 2D). Pour

cela, il faut prendre des covariances assez grandes pour définir les matrices d'erreur. La définition de la matrice d'erreur d'observation $\mathbf{R}$ est plus délicate car nous ne possédons pas de réalité terrain pour mesurer la valeur de l'erreur. On a donc choisi de régler ces covariances en fonction de déplacements virtuels en pixels (empiriquement) dans l'image de la caméra statique en terme d'accélération maximale pour des personnes.

La matrice $\mathbf{H}$ équivaut à une matrice identité de taille 4×4.

5.1.3 Commande de la caméra dynamique

L'objectif de la commande de la caméra dynamique est de se centrer sur l'objet détecté et d'adapter le zoom afin d'avoir une meilleure résolution de la cible.

Détermination des angles

Le calibrage de notre système donne une relation directe entre une position (x, y) dans l'image de la caméra dynamique et les angles à appliquer pour que la caméra dynamique soit centrée sur cette position. Comme la méthode de soustraction de fond utilisée fournit le centre de gravité de la boîte englobant la cible à suivre, on a ainsi, par application directe du résultat de la méthode de calibrage, les angles pour commander la caméra dynamique.

De nombreux travaux en automatique proposent des schémas d'asservissement performants et robustes comme, par exemple, un asservissement basé sur un régulateur RST [62]. Généralement, l'asservissement est basé sur une commande en vitesse ce qui permet d'obtenir un suivi en continu. Malheureusement, le matériel utilisé ne se prête pas à cela (se reporter à l'annexe). D'une part, les requêtes de commande de la caméra dynamique ne sont pas prévues pour effectuer une commande en vitesse. D'autre part, utilisant des caméras réseaux, nous avons aucune connaissance sur le temps de réponse de notre système. En effet, aucune contrainte de temps ne peut être appliquée sur l'exécution de requêtes sur le réseau. Par ailleurs, les caméras réseaux utilisées ne renvoient pas de code de retour indiquant si la requête a été exécutée ou non. Or, les systèmes de régulation sont généralement basés sur un historique de plusieurs observations. Il est alors nécessaire de surestimer le temps de réponse de notre système afin de s'assurer d'englober l'envoi et l'exécution de la requête (de l'ordre de 2s). Dans ces conditions, le temps de réaction de l'asservissement ne répond plus aux contraintes temps réel que l'on s'est fixé.

Nous avons donc choisi de commander la caméra dynamique en position. L'inconvénient majeur est le fait que le suivi se fait en saccade : entre chaque commande, le caméra s'arrête puis, doit repartir et ainsi de suite. Afin de limiter cette impression de saccade et de se rapprocher de la commande en vitesse, on a choisi de commander la caméra à chaque itération du traitement de détection et de suivi effectué dans la caméra statique. Ainsi, la distance parcourue par la cible entre deux images consécutives est faible : le déplacement angulaire est donc petit et limite les inconvénients de la commande en position.

Lorsqu'une cible est immobile dans l'image, le résultat de la soustraction de fond oscille autour de la position optimale. Afin d'éviter de faire osciller la caméra, on contraint la

caméra à se déplacer seulement si la distance entre la position courante et la position estimée est supérieure à une distance minimale prédéfinie.

Détermination du zoom

Le but est d'estimer la commande de zoom à appliquer afin que le champ de vue de la caméra dynamique soit ajusté à la taille de la cible. Autrement dit, on cherche la valeur du zoom telle que la plus grande dimension de la cible occupe $x\%$ de la dimension de l'image correspondante : le facteur x étant prédéfini.

On note $dimX$ et $dimY$ les dimensions de l'image de la caméra dynamique. Dans un premier temps, on calcule les rapports $\frac{dimX}{h_x}$ et $\frac{dimY}{h_y}$, où h_x et h_y sont les dimensions du blob dans I_s. Ensuite, grâce à la table de correspondance apprise par la méthode de calibrage présentée au chapitre 2, on détermine le zoom tel que le plus petit des deux rapports calculés soit ramené à $x\%$.

Idéalement, afin de s'adapter au mieux à la taille de la cible au cours du suivi, la commande de zoom devrait se faire à chaque itération comme la commande de position vue précédemment. La caméra dynamique que l'on utilise ne nous permet pas de commander le zoom à chaque itération : la commande s'exécute sur plusieurs itérations, variables selon la valeur de la commande appliquée. Par conséquent, nous avons choisi d'appliquer la commande de zoom au minimum toutes les n itérations (n étant estimé par expérimentation).

De plus, pour les mêmes raisons que la commande en position, on impose une contrainte supplémentaire à la commande de zoom : la commande est appliquée seulement si la différence entre la taille courante et la taille estimée de la cible est supérieure à un seuil.

5.1.4 Limites du système maître-esclave

Suivi d'une personne seule

On se place dans le cas où une seule personne est présente dans la scène. Durant l'expérimentation, il lui a été demandé de marcher dans un couloir le plus naturellement possible. La figure 5.6 présente le résultat du suivi de personne. On note qu'on suit la personne tout au long du déplacement avec une bonne adaptation du zoom. On arrive donc au même résultat que Bodor *et al.* [13] avec un calibrage faible de la paire de caméra.

Cas problématiques

Certains cas deviennent problématiques au vu de la simplicité de la méthode. En effet, la méthode de soustraction de fond utilisée détecte *tout* objet en mouvement dans la scène sans aucune distinction. Trois cas problématiques à résoudre ont été identifiés :

◇ *Détection multiple* :

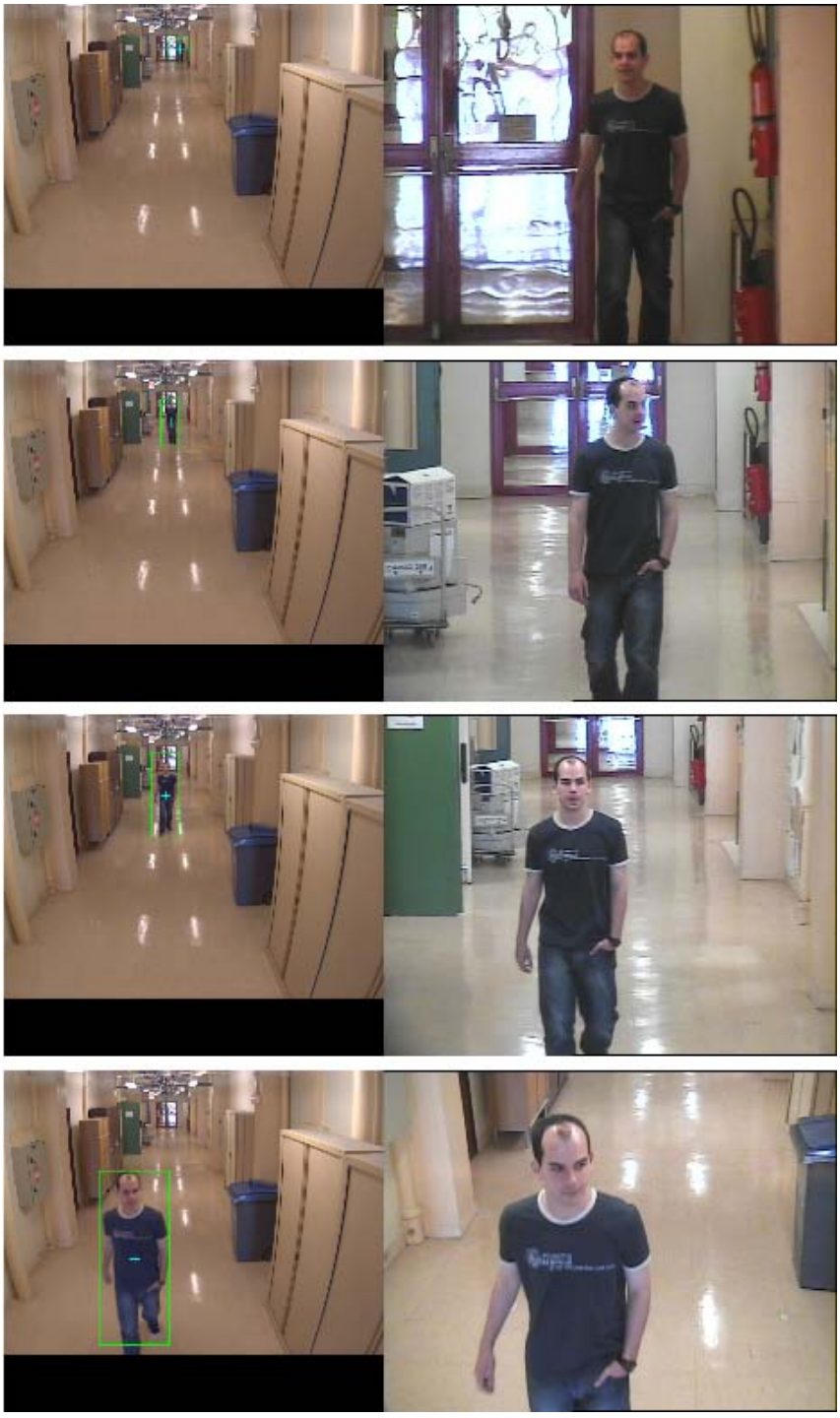


Figure 5.6: *Illustration du suivi maître-esclave d'une seule personne dans un couloir avec adaptation du zoom.*

Dans le cas précédent, on s'est placé dans un environnement contrôlé afin d'avoir des conditions idéales (une seule personne dans une scène vide). En pratique, on rencontre rarement ce cas. En effet, plusieurs personnes peuvent être présentes dans la scène et la soustraction de fond fournit alors plusieurs blobs comme résultat de détection (figure 5.7). Il faut donc à l'instant inital du suivi choisir une cible, puis la suivre. Ce problème peut être résolu par la mise en place d'une méthode de suivi. Ici, on a fait le choix d'utiliser un filtre de Kalman étendu. Les résultats seront présentés dans le chapitre 6.

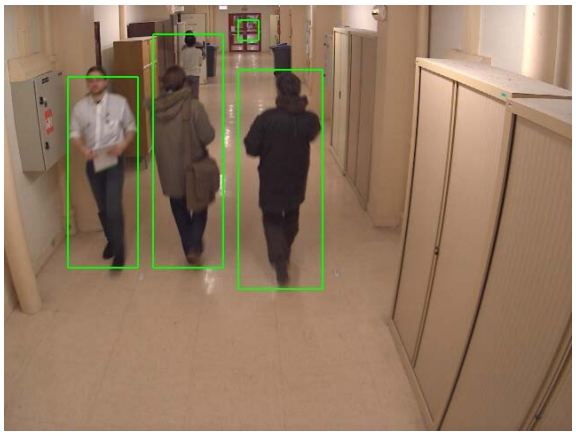


Figure 5.7: *Cas de détection multiple dans la scène.*

◇ *Erreur de détection* :
La soustraction de fond ne permet pas de différencier une personne dont les couleurs de ses vêtements sont ressemblantes à la scène. Dans ce cas, une personne sera décrite de manière partielle par plusieurs blobs, figure 5.8.
Le problème à résoudre dans ce cas est de déterminer de quelle manière raffiner l'information pour que les paramètres de commande envoyés à la caméra dynamique corresponde à une personne entière pour avoir un suivi cohérent au cours du temps.

◇ *Blob contenant plusieurs personnes* :
La méthode de soustraction de fond utilisée est basée sur le traitement de blocs de 8×8 pixels. Ceci a pour conséquence directe que si deux personnes sont trop proches l'une de l'autre dans l'image de la caméra statique (figure 5.9), une connexion entre les deux zones de détection est possible. Dans ce cas de figure, la détection à l'instant t va fusionner les deux détections distinctes à l'instant $t-1$ dans le même blob.
Si on se retrouve dans ce cas lors du suivi de la cible considéré, momentanément, la taille et le centre de gravité du blob ne vont plus correspondre à cette cible. Le problème à résoudre est donc de déterminer de quelle manière raffiner l'information extraite de la soustraction de fond afin que les paramètres de la caméra dynamique soient adaptés à la cible que l'on suit.

Figure 5.8: *Cas de mauvaise détection de cible par la méthode de soustraction de fond.*

Figure 5.9: *Cas de détection dans lequel plusieurs personnes sont dans un même blob (fusion de blob).*

Solution envisagée

Dans le cas des erreurs de détection ou d'un blob contenant plusieurs personnes dont la cible suivie, l'information brute issue de la soustraction de fond ne suffit pas pour fournir une commande cohérente au suivi d'une personne spécifique à la caméra dynamique. Il est nécessaire d'obtenir un supplément d'information afin de pouvoir différencier les différentes personnes présentes dans le blob, donc de caractériser l'apparence des personnes par la couleur, les contours, la texture...

Or, dans l'image de la caméra dynamique, une personne dans le fond de la scène à une hauteur de l'ordre de 40 pixels alors que dans l'image de la caméra dynamique avec le zoom adéquat la taille d'une personne est de l'ordre de 400 pixels. Il semble donc naturellement plus avantageux de chercher une information supplémentaire basée sur la couleur ou la texture pour raffiner la commande dans l'image de la caméra dynamique que dans la caméra statique du fait de la meilleure résolution. Ainsi, cette difficulté peut être résolue en combinant des informations issues de la détection et du suivi de cible dans la caméra statique et d'un autre algorithme de suivi de la même cible dans la caméra dynamique. Nous allons proposer dans la partie suivante un système collaboratif.

5.2 Système collaboratif multi-caméras

Le suivi basé sur un système maître-esclave a l'avantage d'être simple de mise en oeuvre mais possède un certain nombre de faiblesses. C'est pour cela que l'on propose un système basé sur une collaboration plus étroite entre les deux caméras pour proposer un suivi de personne robuste et haute résolution.

Comme on a pu le voir lors du chapitre précédent, les méthodes de fusion de données hétérogènes venant de divers capteurs les plus utilisées dans ce domaine sont les méthodes probabilistes. Du fait de sa plus grande généricité, nous avons choisi d'opter pour une méthode utilisant un filtre à particules. En effet, ce type de méthode possède un certain nombre d'avantages pratiques [87] :

- *facilité de mise en place* : la mise en œuvre algorithmique est extrêmement simple.
- *maintien simple d'hypothèses multiples* : il n'est pas nécessaire, comme dans d'autres techniques de suivi, d'employer un mécanisme complexe de maintien d'hypothèses multiples, l'emploi d'échantillons réalisant cela naturellement. Chaque échantillon est en soi une hypothèse individuelle. Cette capacité à suivre facilement plusieurs hypothèses au cours du temps confère au filtre particulaire une grande robustesse vis-à-vis du fouillis, des changements brusques d'apparence, des occlusions partielles ou totales, ou des accroissements soudain de l'amplitude de mouvements.
- *absence quasi totale de contrainte sur le modèle* : la loi de donnée peut être complexe à souhait tant qu'il est possible de l'évaluer à une constante près. En particulier, comme nous l'avons vu précédemment, les observations peuvent comporter plusieurs modalités (couleur, forme, mouvement, ...), qui sont facilement combinées en une loi produit sous forme d'hypothèse d'indépendance conditionnelle.

Du fait de ces différents avantages et notamment de la simplicité d'intégrer l'information venant d'un suivi effectué dans la caméra statique, on a choisi d'appliquer dans la caméra dynamique une méthode de suivi basée sur un filtre à particules et de fusionner les différentes données lors de l'estimation de la fonction de propagation des particules.

Il reste à définir quel type de modèle d'observation l'on va utiliser pour déterminer le taux de vraisemblance entre le modèle et les candidats. D'après Porikli [83], un suivi basé sur la matrice de covariance de caractéristiques visuelles donne de très bons résultats notamment dans le cas de changement d'apparence et de mouvement irrégulier et rapide. Mais en contre partie, ce type d'approche reste très lent par rapport aux méthodes utilisant les histogrammes couleurs. Comme la contrainte temps réel est primordiale dans notre cas, on va préférer une méthode basée sur les histogrammes couleurs.

De plus, ce type de représentation est utilisée avec succès dans des applications de suivi car il a l'avantage d'être robuste aux changement de pose et d'apparence de l'objet suivi. Cependant, la perte de l'information spatiale réduit la spécificité du modèle et donc augmente les risques de déviation de l'algorithme de suivi vers d'autres objets de la scène. Afin de passer outre cette limitation, Birchfield et Rangarajan ont proposé récemment un nouveau concept, nommé *spatiogramme* [9], vu comme une généralisation de la notion d'histogramme auquel on ajoute une information spatiale, moyenne et covariance pour chaque classe de l'histogramme. Afin d'être plus robuste, notre modèle d'observation sera basé sur l'approche des spatiogrammes couleurs.

Au final, notre système collaboratif multi-capteurs est composé des éléments suivants qui seront détaillés dans la suite du chapitre :

- ◇ *caméra dynamique* : suivi d'objet avec un filtre à particules dont le modèle d'observation est basé sur l'approche spatiogramme,
- ◇ *fusion de données* : intégration des détections de la caméra statique dans la fonction de propagation afin d'optimiser la diffusion des particules dans les zones d'intérêt,
- ◇ *commande de la caméra dynamique* : adaptation de la position et du zoom de la caméra au déplacement de la cible dans la scène.

5.2.1 Filtre à particules *Sequential Important Resampling*

Apparu dans les années 70 mais nécessitant une grande puissance de calcul par rapport aux moyens informatiques de l'époque, les filtres particulaires furent repris de manière indépendante dans les années 90 par Del Moral *et al.* [73], Gordon *et al.* [37], Kitagawa *et al.* [57] et Isard et Blake [48]. Les appellations respectives données par leurs auteurs sont le filtre particulaire avec interaction, le filtre boostrap, le monte-Carlo filter et l'algorithme dit de condensation.

Les méthodes de filtrages particulaires approchent la distribution de probabilité a posteriori $p(x_k|z_{1:k})$ par une mesure discrète égale à une somme finie de mesures de Dirac centrées en des points appelés *particules* et pondérées par des coefficients appelés *poids des*

particules.

$$p(x_k|z_{0:k}) = \sum_{i=0}^{N} \omega_k^i \delta\left(x_k - x_k^i\right) \tag{5.8}$$

où ω_k^i représente le poids de la particule x_k^i, $\delta(.)$ désigne la distribution de Dirac et N le nombre de particules utilisés pour estimer la densité de probabilité a posteriori $p(x_k|z_{1:k})$.

Les particules représentent des hypothèses sur l'état de la cible. Le poids associé à une particule est proportionnel à la probabilité de cette hypothèse. Les méthodes de filtres particulaires consistent à propager la liste des particules pondérées au cours du temps de sorte que les particules explorent l'espace d'état de la cible et se concentrent par un mécanisme de sélection dans les régions de forte ressemblance de l'espace d'état [1].

Parmi les différentes approches de filtres à particules existantes, nous avons choisi d'utiliser le filtre particulaire à échantillonnage pondéré séquentiel avec ré-échantillonnage, *Sampling Important Resampling* en anglais (noté SIR). Le filtre particulaire SIR permet de propager les particules selon une fonction autre que la loi d'évolution. Une telle propagation permet de mieux guider les particules lors de leurs propagations en tenant compte des observations à travers une fonction de proposition.

L'algorithme du filtre à particules SIR permettant d'estimer la densité de probabilité a posteriori $p(x_k|z_{1:k})$ est présenté par la suite. On présente une itération de l'algorithme du filtre à l'instant k en supposant connu les particules et leurs poids à l'instant $k-1$.

Les diverses étapes de l'algorithme sont les suivantes :

- Prédiction :
  Pour $i = 1, \ldots, N$
  Tirage aléatoire des particules $\tilde{x}_k^i$ selon une fonction de de propagation q,
  Calcul du poids des particules tel que $\tilde{\omega}_k^i = p(z_k|x_k^i)$,
- Mise à jour des poids :
  Calcul du poids total : $t = \sum_{i=1}^{N} \tilde{\omega}_k^i$,
  Pour $i = 1, \ldots, N$, normalisation des poids $\omega_k^i = \frac{1}{t}\tilde{\omega}_k^i$,
- Ré-échantillonnage :
  Pour $i = 1, \ldots, N$, tirage avec remise de x_k^i parmi $\left\{\tilde{x}_k^i\right\}_{i=1}^{N}$ proportionnellement aux poids $\left\{\omega_k^i\right\}_{i=1}^{N}$.

5.2.2 Espace d'état

Nous nous intéressons à la position et à la taille de la cible. L'état de la cible est représenté par le plus petit rectangle qui regroupe tous les pixels appartenant à la cible. Ce rectangle est défini par la position (x_r, y_r) de son centre, sa largeur h_x et du rapport Φ entre la largeur h_x et la hauteur h_y du rectangle.

Le vecteur d'état est donc composé de quatre paramètres :

$$X = \{x_r, y_r, h_x, \Phi\}. \tag{5.9}$$

On suppose que la posture de la personne ne varie pas trop au cours du temps c'est-à-dire qu'elle ne s'accroupit pas ni ne s'allonge durant la phase de suivi. On pose donc que le rapport Φ est constant au cours du temps c'est-à-dire :

$$\Phi^0 = \Phi^1 = \cdots = \Phi^N.$$

Cette considération permet de réduire la dimension de l'espace d'échantillonnage aux trois paramètres suivants :

$$X = \{x_r, y_r, h_x\}. \tag{5.10}$$

L'objectif du filtrage particulaire consiste donc à estimer les paramètres du rectangle englobant la cible sur chaque image de la séquence. Une particule du filtre correspond à une hypothèse de rectangle englobant la cible. Même si le rectangle n'est pas toujours adapté à la forme d'une personne, il permet néanmoins de restreindre l'espace d'état et de constituer un bon compromis entre l'adéquation de la représentation de la forme et le nombre de paramètres à estimer. Concernant la description de la dynamique du modèle, nous avons choisi d'utiliser un modèle dynamique d'ordre 0 (vitesse nulle). Le position finale du filtre à particules est obtenue par l'application sur l'ensemble de la population de particules d'un estimateur MMSE correspondant à la moyenne des particules pondérées.

5.2.3 Modèle d'observations : Spatiogramme

Lors de l'initialisation du suivi, nous construisons à partir de la première image de la séquence et plus précisément à partir des pixels contenus dans le rectangle englobant la cible, le spatiogramme modèle de la cible qui constituera notre référence. Ensuite, pour chaque particule obtenue à l'étape de prédiction, on calcule le spatiogramme sur la région délimitée par le rectangle. On obtient ainsi des spatiogrammes candidats. L'objectif de de l'étape de correction est de favoriser les spatiogrammes candidats similaires au spatiogramme de référence h^* en définissant le poids des particules auxquelles se rapportent les spatiogrammes par :

$$\omega_k^i = exp\left(\lambda\rho(h_{x_k^i}, h^*)\right). \tag{5.11}$$

où $\rho(h_{x_k^i}, h^*)$ représente la similitude entre le spatiogramme candidat $h_{x_k^i}$ évaluée sur la région délimitée par le rectangle donné par la particule x_k^i et le spatiogramme de référence h^*. La constante λ représente un coefficient multiplicateur.

Dans les parties qui vont suivre, nous allons définir tout d'abord ce qu'est un spatiogramme puis la notion de similitude entre deux spatiogrammes, c'est-à-dire $\rho(h_{x_k^i}, h^*)$.

Spatiogramme

On considère une image couleur I définie par un ensemble discret de points. L'histogramme de I est un vecteur de valeurs donné par $h_I(b) = n_b,\ b = 1, \ldots, B$ où n_b représente le nombre de point de I ayant la couleur associée à la classe de label b et B le nombre de classes de l'histogramme. Pour une image en niveaux de gris composé de 256 niveaux de

gris, $h_I(b)$ représente le nombre de pixels de l'image tel que la valeur du niveau de gris du pixel vaut b.

Dans [9], Birchfield et Rangarajan considèrent l'histogramme comme un spatiogramme d'ordre zéro. Le spatiogramme d'ordre deux qu'ils proposent est un triplet de valeurs tel que : $h_I(b) = \langle n_b, \mu_b, \Sigma_b \rangle$, $b = 1, \ldots, B$ où μ_b et Σ_b représentent respectivement le vecteur moyenne et la matrice de covariance des coordonnées spatiales des pixels contribuant à la b$^{\text{e}}$ classe.

Afin d'illustrer la différence entre les histogrammes couleur et les spatiogrammes, Birchfield et Rangarajan proposent dans [9] une représentation visuelle du résultat, sur trois images représentant le visage d'une personne, du calcul de l'histogramme et du spatiogramme. Ce résultat est illustré par la figure 5.10. On voit bien que le spatiogramme capture la relation spatiale entre les couleurs composant l'image, alors que l'histogramme perd toute information spatiale.

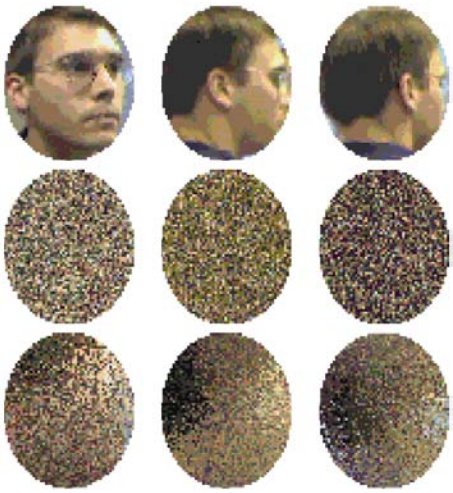


Figure 5.10: *Trois positons différentes d'un visage (première ligne), avec des images générée à partir du calcul de l'histogramme (ligne du milieu) et du spatiogramme de l'image (ligne du bas) [9]. Le spatiogramme capture la relation spatiale entre les couleurs composant l'image, alors que l'histogramme perd toute information spatiale.*

Similarité entre deux spatiogrammes

Le calcul de la similitude entre deux spatiogrammes est équivalent à la somme pondérée de la similitude entre deux histogrammes :

$$\rho(h, h') = \sum_{b=1}^{B} \psi_b \rho_n(n_b, n'_b). \tag{5.12}$$

Nous avons choisi d'utiliser la même formulation du calcul de la similitude $\rho(h, h')$ que

Birchfield et Rangarajan proposée dans [10], formulation plus robuste que celle proposée par O'Conaire *et al.* [27].

Le poids ψ_b est proportionnel à une gaussienne de moyenne μ'_b et de covariance $2\left(\Sigma_b + \Sigma'_b\right)$ évaluée en μ_b :

$$\psi_b = 8\pi|\Sigma_b\Sigma'_b|^{\frac{1}{4}}\mathcal{N}\left(\mu_b; \mu'_b, 2\left(\Sigma_b + \Sigma'_b\right)\right). \tag{5.13}$$

La similitude entre deux histogrammes est ainsi donnée par la formule suivante :

$$\rho_n(n_b, n'_b) = \frac{\min\left(n_b, n'_b\right)}{\sum_{j=1}^{B} n_j}. \tag{5.14}$$

5.2.4 Fonction d'importance

Le filtre à particules SIR a l'avantage d'accepter toute fonction de propagation à l'étape de prédiction. Afin d'affiner la propagation des particules dans les zones d'intérêt, c'est-à-dire en restreignant la région de propagation aux zones où la caméra statique a détecté de l'activité, nous avons choisi de définir une nouvelle fonction de propagation q dépendant de la probabilité $p(x_k|x^i_{k-1})$ mais aussi de la détection d'objets en mouvement par la caméra statique. L'information de la caméra statique est définie comme un mélange de gaussiennes de paramètres μ_j et Σ_j. On obtient ainsi une fonction de propagation q définie de la manière suivante :

$$q\left(p(X_k|X^i_{k-1}), \sum_j \mathcal{N}(\mu_j, \Sigma_j)\right) \tag{5.15}$$

Okuma *et al.* dans [76] et Perez *et al.* dans [88] utilisent une somme pondérée de la probabilité $p(X_k|X^i_{k-1})$ et du mélange de gaussienne décrivant l'information supplémentaire ajouté à la fonction de propagation, schématisée sur la figure 5.11.

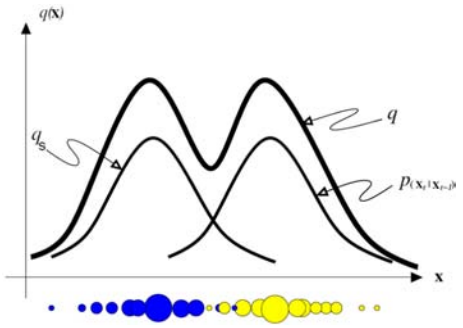


Figure 5.11: *Création de la fonction de propagation q comme la somme de la probabilité $p(x_k|x^i_{k-1})$ et du mélange de gaussienne q_s décrivant l'information venant de la caméra statique [76].*

Ainsi, appliqué à notre cas, on obtient la relation suivante :

$$q(X_k|X_{k-1}^i, z_k) = \alpha\ p(X_k|X_{k-1}^i) + (1-\alpha)\ \sum_j \mathcal{N}(\mu_j, \Sigma_j) \tag{5.16}$$

où α est le coefficient de pondération indiquant le poids que l'on veut donner aux informations supplémentaires.

La propagation des particules, répondant à la relation 5.16, concerne uniquement les paramètres de position (x_r, y_r) des particules. Le paramètre de taille h_x suit simplement une loi gaussienne afin de limiter l'impact des problèmes de détections liés à la caméra statique. Mais afin d'alléger les notations, nous avons gardé dans la relation 5.16 la notation X qui regroupe tous les paramètres.

Il nous reste à déterminer le nombre d'éléments composant le mélange de gaussiennes ainsi que leurs paramètres.

Une manière simple est d'affecter à chaque blob résultant de la soustraction de fond dans la caméra statique une gaussienne 2D dont la moyenne est le centre du blob et la variance est fonction de la taille du blob. Cette approche simpliste est suffisante dans le cas idéal : un blob correspond à un et un seul objet (cas (a) de la figure 5.12). Malheureusement, en pratique, il n'est pas rare que plusieurs personnes soient trop proches pour que la méthode de soustraction de fond puisse les différencier : aux temps $t = 0$ jusqu'au temps t, deux personnes sont détectées séparément par la soustraction de fond et, au temps $t + 1$, l'algorithme ne distingue plus qu'un seul objet en mouvement : il y a fusion de blobs (cas (b) de la figure 5.12). On perd donc de l'information et de la précision.

Afin d'obtenir l'information de fusion de blob pour distinguer les personnes composant le blob pour déterminer un mélange de gaussiennes mieux adapté à la réalité (cas (c) figure 5.12), on va chercher à découper ce blob en un certain nombre de sous-ensembles en fonction du nombre de personnes présentes dans le blob puis d'estimer les paramètres des gaussiennes pour chaque sous-ensemble.

Estimation des paramètres des gausiennes

Supposons que le blob soit découpé en K sous-ensembles. Afin de déterminer les paramètres des K gaussiennes, on utilise un algorithme itératif simplifié appelé E-M (*Expectation-Maximization* en anglais) composé de deux étapes :

1. **Estimation :** calcul de l'espérance de la vraisemblance en tenant compte des dernières variables observées.
2. **Maximisation :** estimation du maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape précédente.

Cet algorithme suppose connu le nombre de gaussiennes. Il faut donc un traitement supplémentaire afin d'estimer en combien de sous-ensembles le blob peut être décomposé.

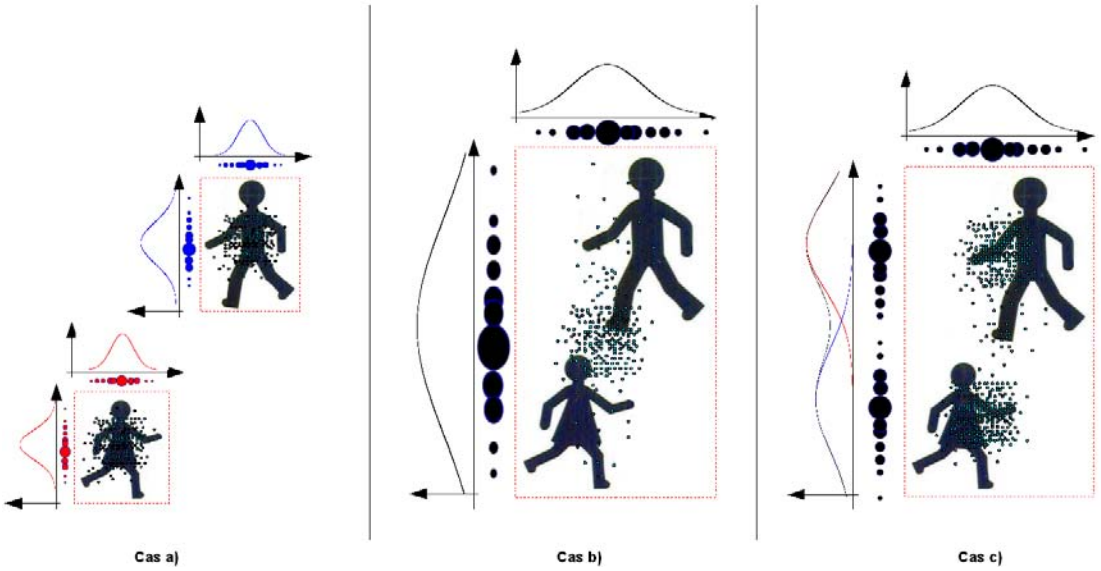


FIGURE 5.12: *Schéma représentant l'estimation du mélange de gaussiennes en fonction des détections issues de la caméra statique. Pour les cas (a) et (b), une gaussienne est calculée par blob qu'il y ait une personne (a) ou plusieurs personnes (b) dans le blob. Dans le cas (c), l'estimation des paramètres de la gaussienne prend en compte la présence de plusieurs cibles dans le blob.*

Méthode de partitionnement de données

En analyse de données statistiques, le partitionnement de données créé peut être dur (chaque exemple est associé à seulement un cluster) ou mou (chaque exemple est associé de manière probabiliste à tous les clusters à des degrés divers).

L'algorithme K-Means ou algorithme des nuées dynamiques est l'algorithme de clustering « dur » le plus connu et le plus utilisé, tout en étant très efficace et simple. C'est un algorithme de clustering non supervisé visant à former une partition optimale en K clusters d'un ensemble de prototypes X. Soit E l'ensemble des prototypes, avec $X_i \in R^n$. Soit une partition de E définissant les K sous-ensembles C_k de E tels que : $\bigcup_{k=1}^{K} C_k = E$; $\forall k \neq j, C_k \cap C_j = \varnothing$ et $\forall k, C_k \neq \varnothing$. Soit f une fonction de coût d'appartenance d'un point X à un classe C_i à valeurs dans R^+ : $f(X, C_i) \in R^+$. L'algorithme K-Means consiste à déterminer par un processus itératif la partition qui minimise : $J = \sum_{K=1}^{K} \sum_{X_i \in C_k} f(X, C_i)$. Pour représenter la classe C_i, on s'appuie sur son noyau. Cela peut être le centre de gravité de la classe, un ensemble de p points représentatifs, les premiers axes d'inertie ou des points tirés au hasard.

Une difficulté avec beaucoup de ces algorithmes est la spécification a priori du nombre de clusters cherchés. De plus, la méthode des K-Means est sensible à l'initialisation. On peut estimer le nombre de blobs fusionnant en un seul blob en mettant en place un suivi de cible en associant une piste à un blob. Ainsi le nombre de pistes associées au blob servira d'initialisation pour la méthode des K-Means : K vaut le nombre de pistes. Par exemple,

si deux blobs fusionnent, deux pistes vont être attribuées au blob résultant, et ainsi, K prend la valeur 2.

Pour cela, on va utiliser le module mis en place pour le suivi de blob dans le cas du système maître-esclave. Pour chaque détection issue de la soustraction de fond, nous appliquons un filtre étendu de Kalman. Une étape supplémentaire d'association de blob aux pistes est ajoutée afin d'évaluer le nombre de pistes qui vont se retrouver associés à un blob.

Schéma intermédiaire de notre système

La figure 5.13 illustre le système mis en place incluant les différents modules présentés précédemment. On note qu'il manque une étape permettant de passer les informations du repère $\mathcal{R}_{I_s}$ aux repère $\mathcal{R}_{I_d}$. Ceci inclut un changement de repère et une restriction des données au champ de vue effectif de la caméra dynamique.

Ainsi le changement de repère des paramètres (μ, Σ), liés respectivement à une position et à une taille d'un blob, est donné par les relations suivantes :

$$\begin{pmatrix} x_d \\ y_d \end{pmatrix} = \mathrm{f}_{s\to d} \begin{pmatrix} x_s \\ y_s \end{pmatrix} \tag{5.17}$$

$$\Sigma_d = \begin{pmatrix} \left(x_d - \mathrm{f}_{s\to d}\left(x_s + \sigma_s^x\right)\right)^2 & 0 \\ 0 & \left(y_d - \mathrm{f}_{s\to d}\left(y_s + \sigma_s^y\right)\right)^2 \end{pmatrix} \tag{5.18}$$

où $\mathrm{f}_{s\to d}$ est issue du calibrage présenté dans le chapitre 2, σ_s^x et σ_s^y sont fonction de la taille du blob.

Le mélange de gaussiennes est composé de B^s gaussiennes réparties sur l'ensemble de l'image de la caméra statique. Or, le champ de vue de la caméra dynamique est plus restreint que celui de la caméra statique. Donc le mélange de gaussiennes défini dans I_s n'est pas défini entièrement dans I_d. On va donc appliquer une fenêtre sur le mélange de gaussiennes afin de ne garder que la partie qui est définie dans I_d.

5.2.5 Commande de la caméra dynamique

On souhaite suivre une cible avec la caméra dynamique telle que le champ de vue de cette dernière soit ajusté à la taille de la cible. Par conséquent, le champ de vue de la caméra dynamique étant restreint, très vite, du fait de son déplacement dans la scène, la cible peut sortir du champ de vue. Afin de garder une image entière de la personne, il faut donc adapter la position ainsi que le zoom de la caméra au déplacement de la cible, ce qui va avoir une incidence sur les paramètres des particules.

Estimation des paramètres de commande

Paramètres angulaires

Grâce à l'étalonnage de la caméra statique, on a directement la relation entre une distance

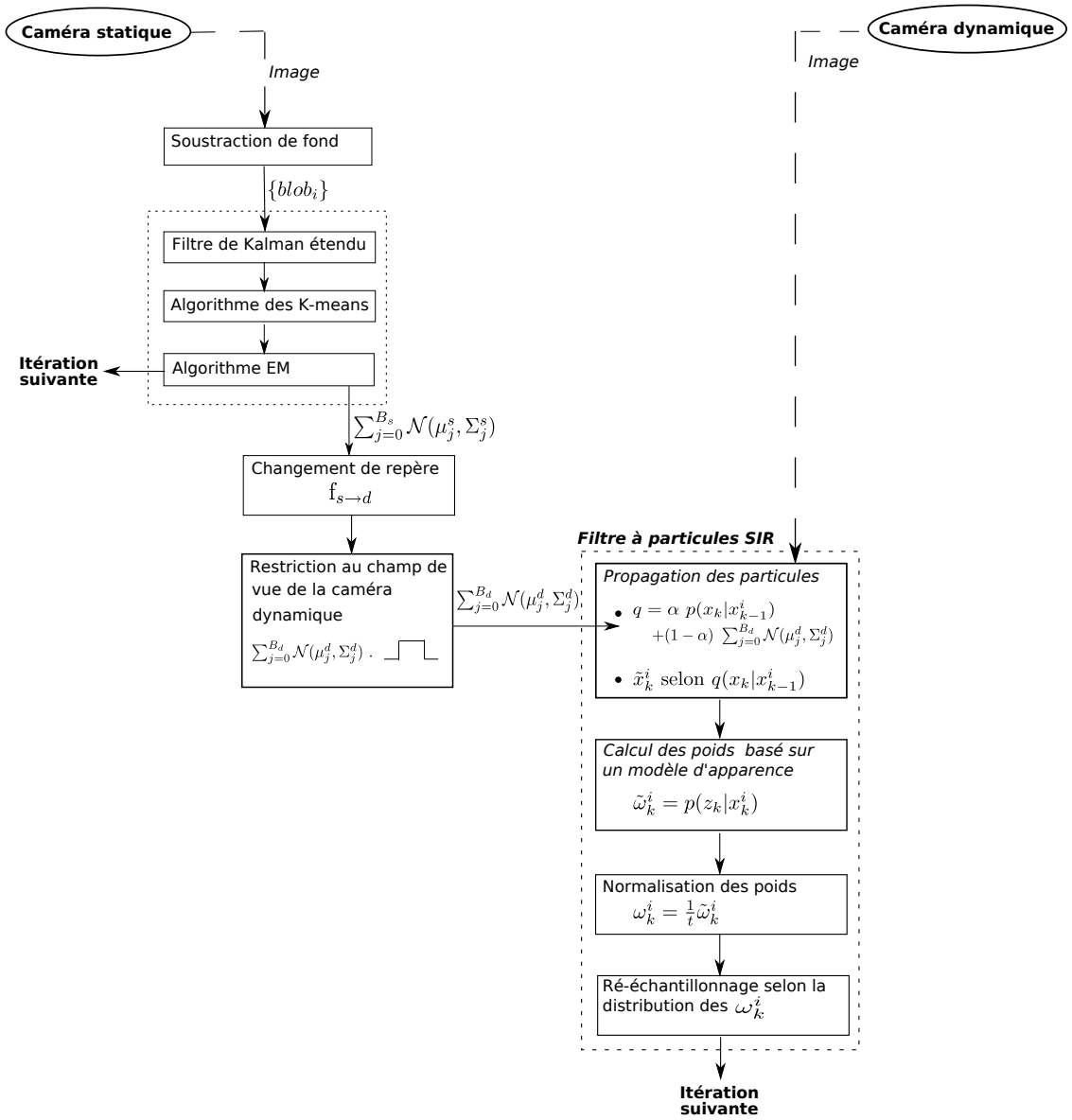


FIGURE 5.13: *Schéma intermédiaire de la collaboration entre la caméra statique et la caméra dynamique : on présente une itération du processus.*

$(\Delta d_x, \Delta d_y)$ en pixels dans I_d et le déplacement angulaire $(\Delta\alpha, \Delta\beta)$ à appliquer aux paramètres angulaires courants de la caméra pour s'adapter au déplacement de la cible. Le déplacement dans l'image $(\Delta d_x, \Delta d_y)$ est déterminé en fonction de la différence entre la position du centre de la cible à l'instant $k-1$ et du résultat du filtre à particules à l'instant k.

Lorsqu'une cible est immobile dans l'image, le résultat du filtre à particules peut osciller autour de la positon optimale. Afin d'éviter de faire osciller la caméra, on contraint la caméra à se déplacer seulement si la distance entre la position courante et la position finale estimée est supérieure à une distance minimale donnée.

De même que lors du suivi selon un processus maître-esclave, on commande la caméra dynamique en position et non en vitesse à chaque itération du filtre. Ce choix suppose que le déplacement de la cible entre l'instant k et $k+1$ est faible ce qui facilite ainsi la convergence du filtre à particules.

Lorsque l'on commande la caméra, on applique une transformation à l'image précédente. Nous faisons l'approximation que chaque rotation de la caméra se traduit dans l'image par une translation dans la direction de la rotation. Il faut donc reporter cette transformation aux paramètres des particules. La transformation consiste en deux translations t_x et t_y, respectivement selon l'axe horizontal et vertical en fonction de la commande appliquée : $t_x = \Delta d_x$ et $t_y = \Delta d_y$. Les paramètres des particules deviennent ainsi :

$$X_{new} = \{x_r + t_x, y_r + t_y, h_x\}\,. \tag{5.19}$$

Paramètre de zoom
L'estimation de la commande de zoom est la même que celle mise en place pour un système maître-esclave. A la différence de la commande en position, on ne reporte pas ce changement d'échelle sur les particules du filtre. Dans le cas idéal où l'application de la commande de zoom serait immédiate, il serait nécessaire d'appliquer le changement d'échelle aux particules immédiatement afin d'adapter au mieux les hypothèses du filtre aux données réelles. Dans notre cas, l'application de la commande de zoom est lente et se fait sur plusieurs itérations du filtre à particules. Comme le changement de taille des données réelles est progressif, on s'aperçoit, après expérimentation, que le filtre à particules est plus performant si on le laisse s'adapter seul à la variation d'échelle progressive que si on force la mise à jour de l'échelle.

Gestion du flou

A cause du mouvement rapide de la caméra dynamique, certaines images présentent un flou de bougé (images 1 et 2 de la première ligne de la figure 5.14) telles que les images ne sont pas exploitables par l'algorithme de suivi. De même, les images de zones homogènes où il n'y a pas d'objet présent (dernière image de la première ligne de la figure 5.14) posent problème. Afin de limiter la divergence du filtre, on va chercher un traitement qui nous permet de détecter ces images afin de ne pas les traiter.

Classiquement, l'effet de flou dans une image est corrigé à l'aide d'algorithmes de restauration d'image basés sur la déconvolution. Dans notre cas, nous ne souhaitons pas restaurer l'image, mais simplement évaluer une valeur qui, comparée à un seuil prédéfini, labellisera l'image comme floue ou non.

Dans [68], Marichal *et al.* proposent une mesure simple mais robuste d'estimation de la qualité d'une image en terme de flou. Leur méthode est basée sur le calcul d'un histogramme des coefficients non nuls de la transformée en cosinus discrète appliquée à l'image. La transformée en cosinus discrète est une fonction mathématique qui permet de changer le domaine de représentation d'un signal. Un signal temporel ou spatial peut être défini dans un espace fréquentiel. Ainsi, une image naturelle flou aura une grande proportion de coefficients nuls du fait de la prépondérance des basses fréquences sur les hautes fréquences.

Tong *et al.* proposent une méthode plus robuste de détection de flou dans [101], utilisant des ondelettes, mais celle-ci est plus lente. Pour des raisons de temps de calcul, nous avons choisi d'implémenter l'approche de Marichal *et al.* qui donne de bons résultats dans le cadre de notre application.

L'algorithme proposé fournit un pourcentage évaluant la part de flou dans l'image. Dans notre cas, une image totalement floue correspond à un pourcentage proche de 100%. Une image nette présente un pourcentage faible. En pratique, après nos expérimentations, nous avons fixé un seuil à 90%, au delà duquel l'image est considérée comme floue et ne peut donc pas être traitée. Les résulats de détection avec un seuil à 90% sont présentés sur la figure 5.14 : la première ligne est classée comme image floue, la seconde comme image nette.

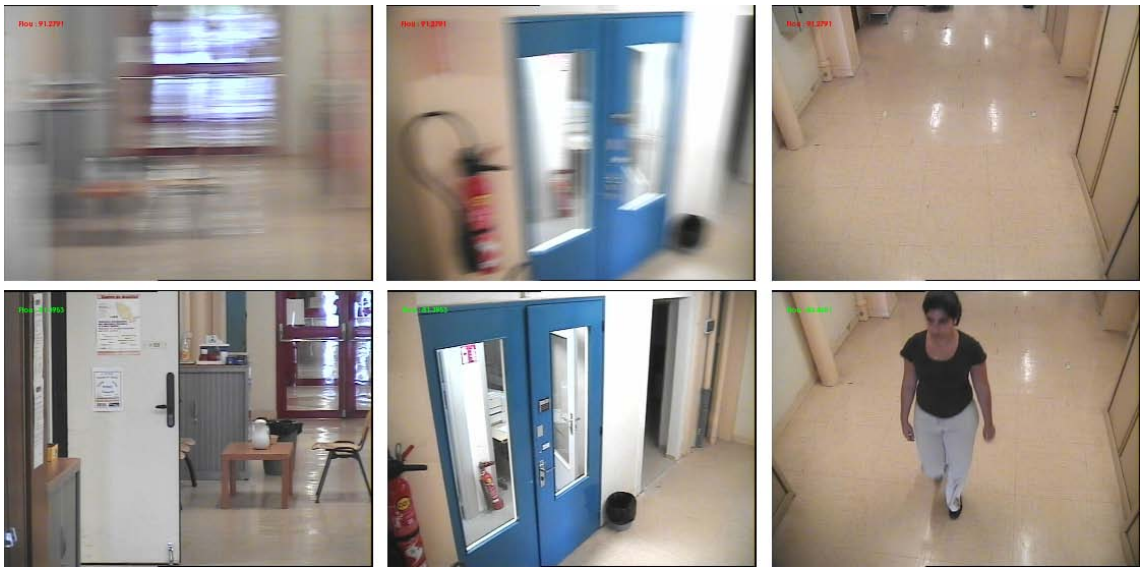


Figure 5.14: *Exemple d'images issues de la caméra dynamique détectées comme floues (première ligne) ou non floues (seconde ligne) donc pouvant être traitées par le filtre à particules. Les pourcentages obtenus sont pour les images « floues »de l'ordre de 91% et pour les autres de l'ordre de 80-85%.*

Schéma final de notre système collaboratif multi-caméras de suivi de personne

La figure 5.15 propose le schéma complet de la collaboration mise en place entre la caméra statique et la caméra dynamique. Les résultats obtenus avec ce système dans le cadre d'un suivi de personne sont présentés dans le chapitre suivant.

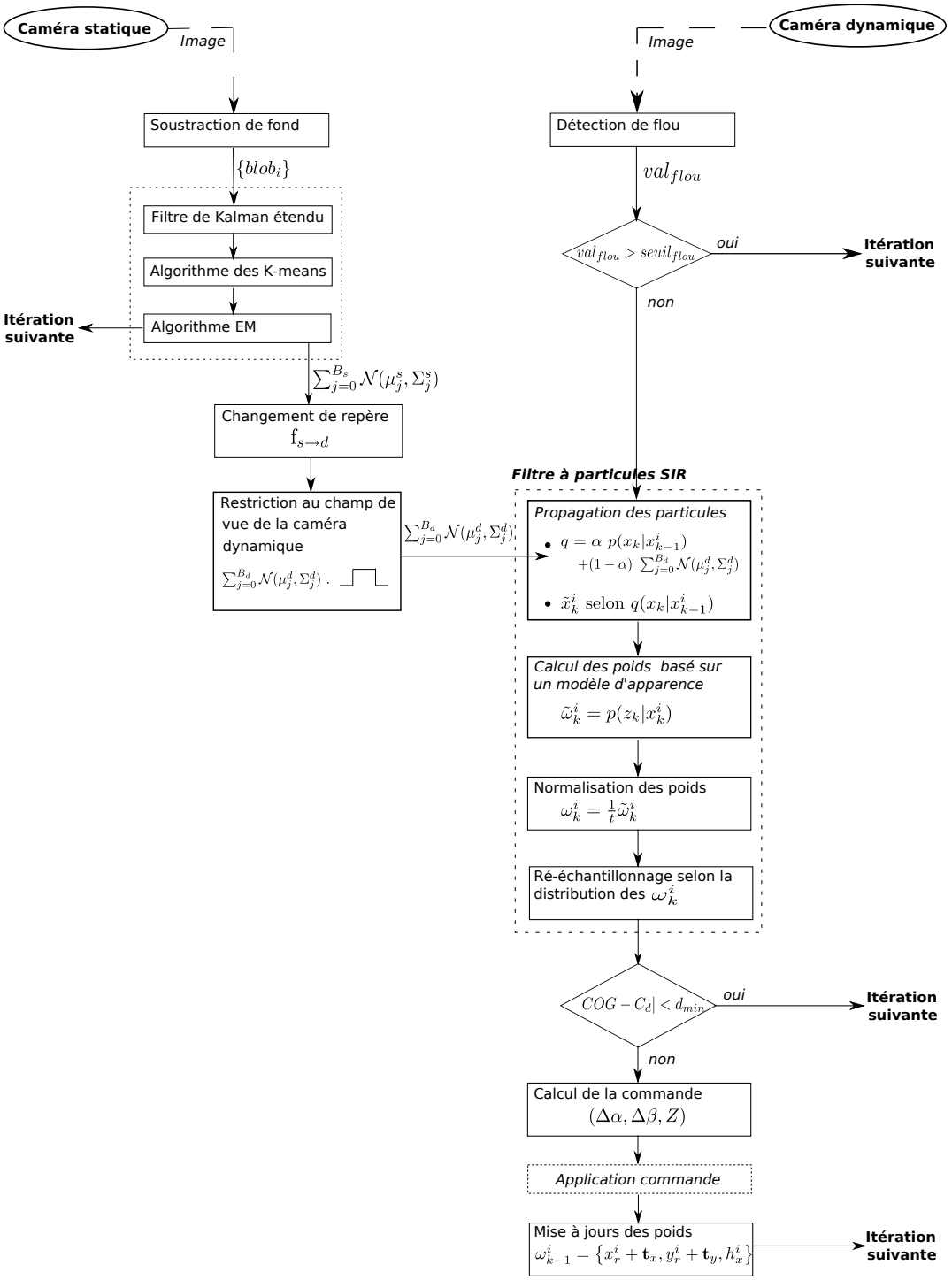


FIGURE 5.15: *Schéma complet de la collaboration mise en place entre la caméra statique et la caméra dynamique : on présente une itération du processus.*

Chapitre 6

Mise en œuvre du système de vision hybride pour le suivi de personne : résultats préliminaires

Dans le chapitre 5, nous avons proposé, dans un premier temps, un processus de suivi utilisant notre système de vision hybride basé sur un schéma maître-esclave. Cette approche ayant un certain nombre de limitations, nous avons présenté, par la suite, un système de suivi de personne basé sur une collaboration plus étroite entre les deux capteurs, fondée sur la fusion de données issues de différentes sources dans le formalisme du filtre à particules.

Dans ce chapitre, un certain nombre de résultats préliminaires vont être présentés. Ces résultats qualitatifs vont permettre de montrer l'apport de la collaboration entre les deux capteurs par rapport à un système maître-esclave notamment dans le cas d'un déplacement rapide de la personne et d'un regroupement de personnes dans la scène.

Les résultats des traitements effectués pour chaque capteur seront présentés de la manière suivante dans les images : les détections d'objets en mouvement par la caméra statique sont matérialisées par une boîte englobante (verte dans les images). Dans le cas de la caméra dynamique, le cadre englobant (magenta dans les images) représente le résultat de l'algorithme du filtre à particules. Les croix de couleur représentent la position spatiale des échantillons avant l'étape de ré-échantillonnage. La couleur est associée au poids de chaque particule sur une palette arc-en-ciel : plus la couleur est proche du rouge et plus le modèle d'apparence lié à la particule est similaire au modèle référence de la cible.

Actuellement, l'initialisation des algorithmes, notamment le choix de la cible et la définition de son modèle d'apparence pour le filtre à particules, est manuelle.

L'évaluation des résultats présentés par la suite est difficile car nous ne pouvons pas traiter a posteriori une même séquence vidéo enregistrée avec les différentes approches (système maître-esclave et système collaboratif) du fait du contrôle en direct de la caméra dynamique. L'évaluation et la comparaison des différents algorithmes s'effectuent donc sur différentes séquences, les plus semblables possibles. Pour cela, on demande aux personnes

faisant partie de l'expérimentation de reproduire au mieux le même scénario pour chaque approche.

6.1 Résultats de suivi lors d'un déplacement rapide de la cible

Pour ce test, il a été demandé à la personne de traverser rapidement le couloir, tout en gardant une démarche naturelle ce qui va impliquer, selon le temps d'exécution de chaque algorithme, une variation importante de la position de la cible entre deux images consécutives.

Dans un premier temps, la personne a été suivie avec un système maître-esclave (figure 5.2). La détection de la personne en mouvement s'effectue dans l'image de la caméra statique. Puis, une commande de position et de zoom est estimée afin d'asservir la caméra dynamique sur la cible. Dans cette approche, les images haute résolution de la caméra dynamique ne sont pas exploitées. Dans un second temps, l'approche de suivi d'une personne basée sur la collaboration entre les deux capteurs est testée.

6.1.1 Suivi dans la caméra statique

La méthode de calibrage utilisée suppose la scène plane. Comme nous l'avons vu au chapitre 3, malgré le fait que la personne suivie ne fasse pas partie de la géométrie de la scène 3D apprise et donc que l'hypothèse de scène plane ne soit plus vérifiée, la caméra dynamique est bien centrée sur la personne, comme on le voit sur l'image n°174 de la figure 6.1. Lorsque la personne est en mouvement, on remarque un décalage entre la position de la personne et le centre de l'image alors que la commande estimée doit permettre de centrer la caméra dynamique sur la cible : à l'image n°188, la cible est à environ 190 pixels du centre de l'image. Ceci est dû d'une part au retard du système entre l'évaluation de la commande à partir de l'information de détection issue de la caméra statique et de son application effective et, d'autre part, à une vitesse apparente lente de la personne dans la caméra statique du fait de la basse résolution dans celle-ci. De plus, la vitesse estimée n'est pas véritablement adaptée à celle de la cible. En effet, du bruit est apporté par le manque de précision de la détection de la cible dans l'image de la caméra statique : le blob est plus ou moins bien adapté à la taille de la cible (image n°188 de la figure 6.1).

Concrètement, on constate qu'entre l'image n°174 et l'image n°177, la personne avance suffisamment dans l'image de la caméra dynamique pour qu'on s'attende à un mouvement de la caméra dynamique pour qu'elle se centre sur la cible. Or, cette commande n'apparaît qu'à l'image n°180. Du côté de la caméra statique, le mouvement entre les deux images est faible, de l'ordre d'une dizaine de pixels et entre l'image n°174 et l'image n°180, la différence est de l'ordre de 30 pixels. Afin d'éviter des oscillations de la commande dues à une instabilité de la détection d'objet, on a contraint le système à s'appliquer la commande que si l'écart entre deux positions consécutives du centre de la cible est supérieure à un seuil. Ainsi, l'écart entre les images n°174 et n°177 est considéré comme un mouvement dû à l'instabilité de l'algorithme de détection alors que celui entre les images n°174 et

n°180 est considéré comme un vrai mouvement de la personne qu'il faut donc corriger. Le seuil est identique quelque soit la position de la personne dans la scène. Or, un même déplacement de la cible semblera plus lent au fond du couloir que proche des caméras : une amélioration possible serait d'adapter la valeur du seuil en fonction de la position de la cible dans l'image.

Un système en maître-esclave permet donc de suivre un déplacement rapide d'une personne grâce à la vision globale de la scène. Cependant, des vitesses rapides peuvent induire des retards importants dans la commande de la caméra dynamique.

6.1.2 Suivi dans la caméra dynamique

Le processus collaboratif proposé est basé sur l'utilisation d'un filtre à particules dans l'image de la caméra dynamique. Comme le montre la figure 6.2, l'utilisation seule d'un filtre à particules ne permet pas de suivre une personne se déplaçant rapidement [83]. On constate que rapidement, la personne sort de la zone où sont distribuées les particules et le filtre à particules ne peut donc plus suivre la cible. Une solution serait d'augmenter l'écart-type de la loi gaussienne sur la dynamique afin que les particules soient distribuées sur une plus grande zone, pour tenir compte de grands déplacements apparents dans l'image de la caméra dynamique. Mais si l'on veut garder une bonne précision, il faut aussi augmenter le nombre de particules. Dans ce cas, le temps de calcul augmente et il n'est plus possible de suivre une personne sur flux vidéo.

De plus, dans le cas où le suivi est effectué uniquement dans la caméra dynamique, si la personne sort du champ de vue de la caméra dynamique, à moins de balayer toute la scène et de rechercher la cible, aucune information n'est disponible pour repositionner la caméra dynamique sur la personne.

6.1.3 Suivi collaboratif

Afin de pallier cette faiblesse du filtre à particules, dans le cas où le suivi est effectué uniquement dans la caméra dynamique, nous avons proposé un processus de collaboration fondé sur la fusion de données venant de chaque caméra. Cela consiste concrètement par la définition d'une nouvelle fonction d'importance q telle que la distribution des particules soit fonction du résultat du filtre à particules sur la caméra dynamique de l'instant précédent et des détections d'objet en mouvement issues de la caméra statique (équation 5.16). Ainsi, comme le montre la figure 6.3, les particules sont diffusées dans les zones de détections et donc peuvent couvrir des déplacements plus importants que dans le cas où le filtre à particules est utilisé seul (figure 6.2), ce qui permet d'avoir des particules qui répondent positivement au niveau de la personne (rouge, orange, jaune). Les particules sont décalées par rapport à la cible car elles sont affichées avant l'étape de ré-échantillonnage.

Actuellement, notre processus de collaboration ne nous permet pas de gérer le cas où la cible sort du champ de vue de la caméra dynamique. Mais, contrairement au cas précédent, grâce au suivi de blobs effectué en parallèle dans la caméra statique, on peut envisager

l'estimation d'un score de confiance qui soit fonction de la cohérence du résultat du suivi collaboratif et de la correspondance entre le champ de vue de la caméra dynamique et la position du blob correspondant à la cible dans l'image de la caméra statique grâce au calibrage. Ainsi, si il n'y a pas de cohérence entre le champ de vue de la caméra dynamique et la position de la cible dans la caméra statique, on considère qu'on ne peut faire confiance qu'à l'information issue de la caméra statique. On bascule alors sur un suivi basé sur le système maître-esclave le temps de recentrer la caméra dynamique sur la personne.

Par rapport au système maître-esclave, le décalage entre le centre de la cible et le centre de l'image est de l'ordre d'une centaine de pixels. Ce décalage est dû seulement au retard induit par le système. De plus, l'adaptation du zoom à la taille de la cible est sensiblement meilleure dans le cas du système collaboratif que dans celui du maître-esclave : dans le premier cas, on adapte le zoom en fonction du modèle caractérisant la personne alors que dans le second cas, il est adapté en fonction de la boîte de détection de zone en mouvement, qui peut contenir une partie de la scène environnante (image n°186 de la figure 6.1). Ce modèle d'apparence haute résolution utilisée dans la caméra dynamique permet donc de mieux estimer la taille de la cible.

6.2 Résultats de suivi d'une personne dans un groupe

Le scénario d'une personne se déplaçant seule dans une scène est simpliste et peu réaliste. En effet, il est courant que plusieurs personnes se trouvent dans un même lieu. On va donc s'intéresser au cas où plusieurs personnes sont présentes dans le scène et peuvent être détectées comme une seule entité par l'algorithme de détection de zones en mouvement dans la caméra statique : cas des images 2 et 3 de la figure 6.4.

Les performances d'un système de suivi fonctionnant en maître-esclave sont fortement contraintes par la qualité de la détection dans la caméra statique. En effet, le zoom est fonction de la taille du blob englobant la cible. Dans le cas idéal, première image de la figure 6.4, une seule personne est présente dans le blob et ainsi, le zoom est parfaitement adapté à la taille de la personne telle que la cible représente 70% de la hauteur de l'image. Dans le cas illustré par les images 2 et 3 de la figure 6.4, le zoom n'est pas adapté à la taille de la cible suivie : la cible représente respectivement 52% et 40% de la hauteur de l'image car le blob est adapté à la zone regroupant plusieurs personnes. Or, du fait de la basse résolution de l'image de la caméra statique, il n'est pas aisé de rajouter un traitement sur l'image de la caméra statique permettant de raffiner la détection afin de choisir parmi toutes les personnes présentes celle que l'on suit.

Notre méthode de collaboration permet d'apporter une solution à ce problème en s'appuyant sur l'information des deux caméras. Dans ce cas, c'est l'utilisation du modèle d'apparence du filtre à particules dans la caméra dynamique qui va permettre de différencier la cible de son environnement grâce à la caractérisation de la cible par un modèle. Ainsi, à part divergence de la méthode, on constate que le champ de vue de la caméra dynamique reste bien adapté à la cible malgré des perturbations extérieures : personnes dans l'environnement de la personne suivie ou détection incomplète (seule le haut de la personne est


image n°174


image n°177

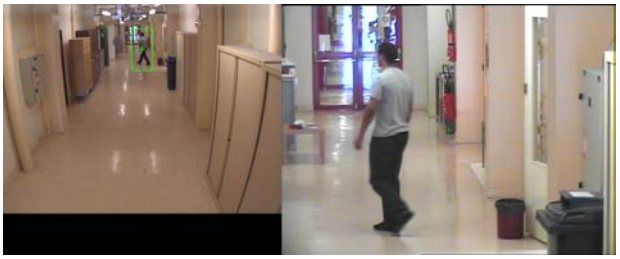


image n°180


image n°183

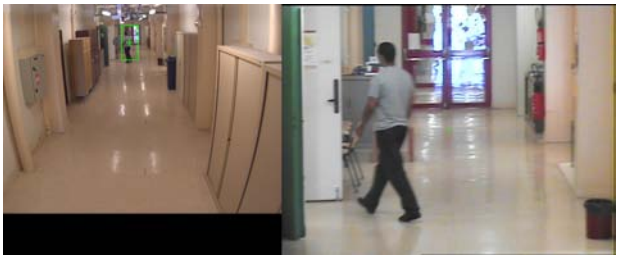


image n°186

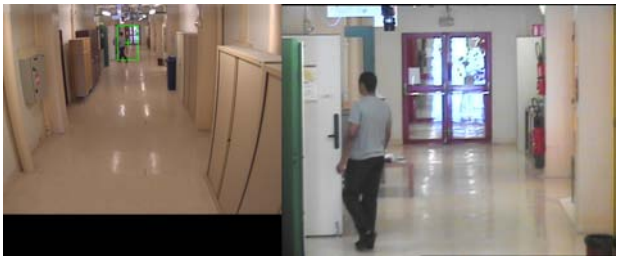


image n°188

FIGURE 6.1: *Suivi d'une personne se déplaçant rapidement dans un couloir dans le cas d'un système maître-esclave.*

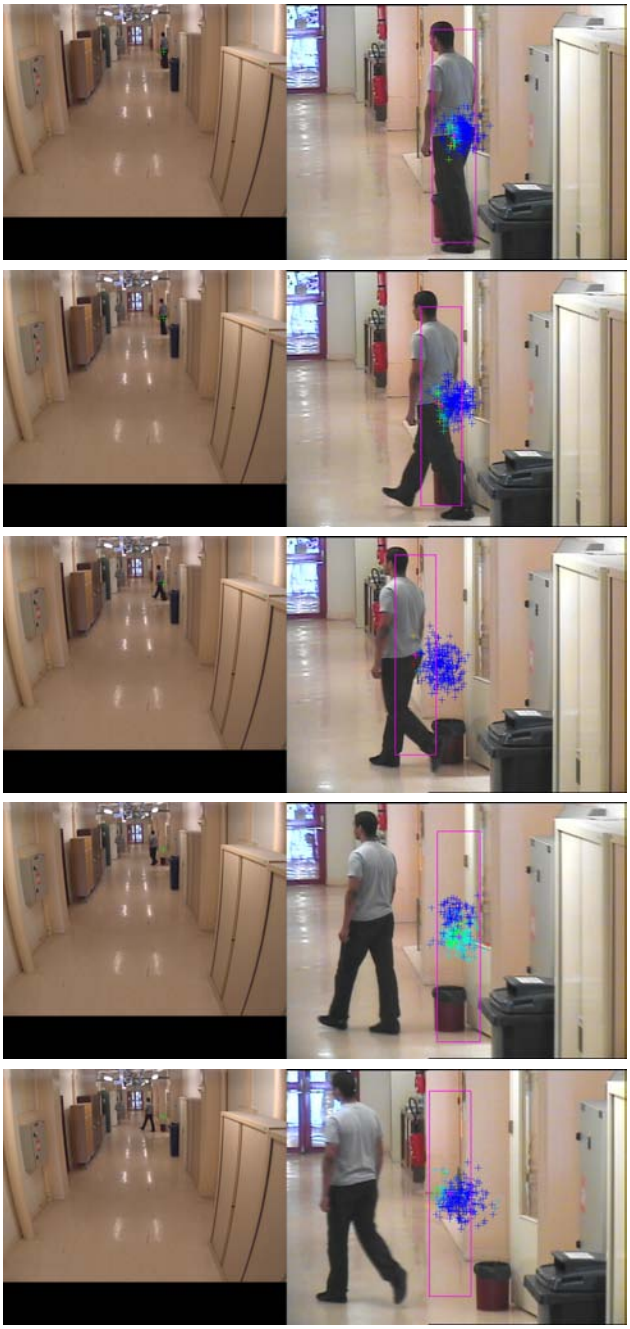


Figure 6.2: *Suivi dans une caméra dynamique (à droite) d'une personne se déplaçant rapidement en utilisant* uniquement *une méthode de filtre à particules.*


image n°32

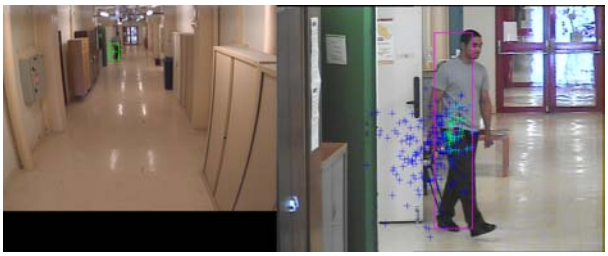


image n°33

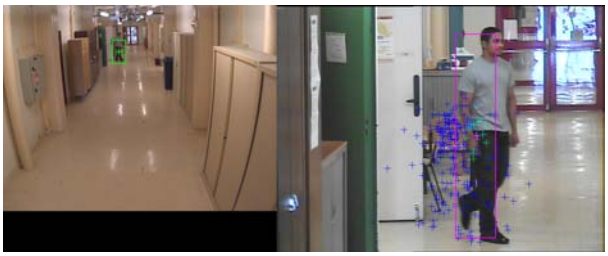


image n°34


image n°36


image n°38


image n°40

Figure 6.3: *Suivi d'une personne se déplaçant rapidement avec notre système de collaboration entre la caméra statique (à gauche) et la caméra dynamique (à droite).*

Figure 6.4: *Résultat du suivi de personne avec le système maître-esclave : cas où il y a une personne pour un blob et cas où plusieurs personnes sont présentes dans un même blob.*

détectée) issue de la caméra statique (figure 6.5), compensée par la caméra dynamique.

6.3 Résultats de suivi collaboratif pour des scénarios plus complexes

Dans cette partie, nous présentons des résultats de suivi haute résolution d'une personne pour un scénario plus complexe que précédemment. Les séquences d'images proposées par les figures 6.6 et 6.7 présentent le suivi haute résolution d'une cible dans un environnement où d'autres personnes sont présentes dans la scène mais sans occultation.

Contrairement aux résultats précédents, la cible reste parfaitement au centre de l'image. En effet, afin de mieux appréhender l'interaction de l'environnement avec la cible suivie, il a été demandé à la personne suivie de marcher lentement afin que le système s'asservisse parfaitement sur elle.

Comme on a pu le voir précédemment, la collaboration entre les deux caméras permet d'effectuer un suivi haute résolution d'une personne malgré la détection imparfaite de la cible (le pantalon de la cible étant de couleur similaire au sol) ou que plusieurs personnes soient considérées comme une seule entité par le module de détection de la caméra statique.

On note aussi sur ces deux séquences d'images que les particules suivent bien la fonction d'importance q définie. En effet, lorsque seule la cible est détectée, les particules sont concentrées sur la cible, comme on le constate sur les images n°94, n°100 et n°111 de la figure 6.6 et n°121 de la figure 6.7. Lorsque plusieurs détections issues de la caméra statique sont contenues dans le champ de vue de la caméra dynamique, les particules sont bien réparties sur les différentes détections ainsi que la cible suivie, comme on le voit sur l'image n°63 de la figure 6.6 et sur les images n°126, n°130 et n°133 de la figure 6.7.

La figure 6.8 illustre le cas où il y a une occultation partielle de la personne suivie. On constate que le filtre à particule est suffisamment performant pour ne pas diverger lors de l'occultation partielle de la cible notamment à l'image n°196 de la figure 6.8.

6.4 Performances et limitations du système de suivi collaboratif

Les premiers résultats de suivi de personne obtenus avec notre système de vision basé sur une collaboration étroite entre une caméra statique et une caméra dynamique sont encourageants. Mais avant d'avoir un système de suivi haute résolution robuste, un certain nombre de points sont à améliorer :

- ◇ *Temps d'exécution* : Nous avons mis en œuvre un schéma algorithmique tirant parti des propriétés de chaque caméra et composé de briques de détection et de suivi efficace pouvant tournant en temps réel. Sur un PC classique avec un processeur Intel Xeon à 3GHz et 3Go de RAM, le traitement se fait en moyenne en 130ms (environ 7 images/s). Ce temps d'exécution est suffisant pour tester le système mais encore trop lent pour

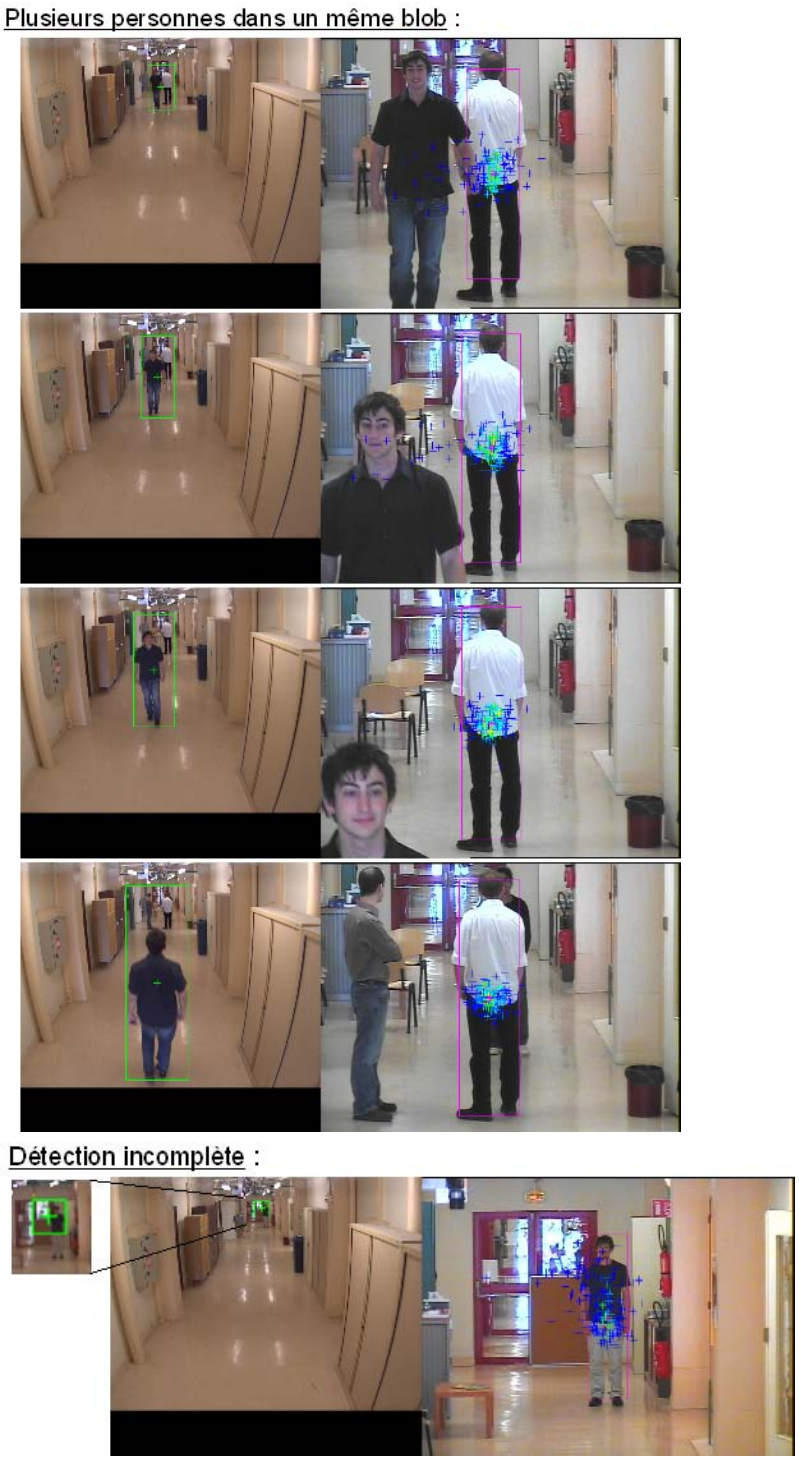


Figure 6.5: *Résultat du suivi de personne basé sur la collaboration entre les capteurs : cas où plusieurs personnes sont présentes dans un même blob et cas où il y a une détection incomplète de la personne dans l'image de la caméra statique (à gauche), compensée par la caméra dynamique.*

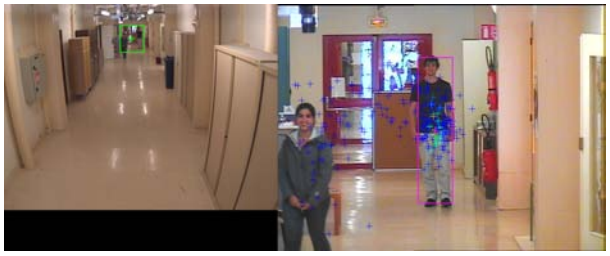


image n°63

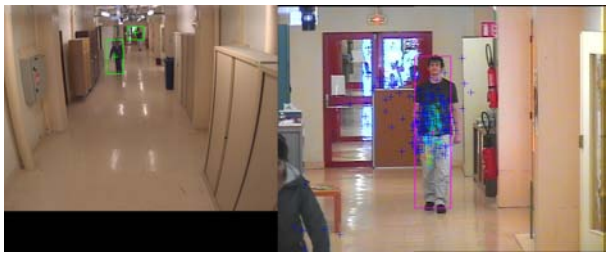


image n°70

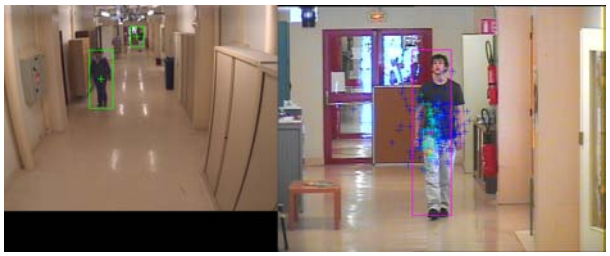


image n°77


image n°94


image n°100

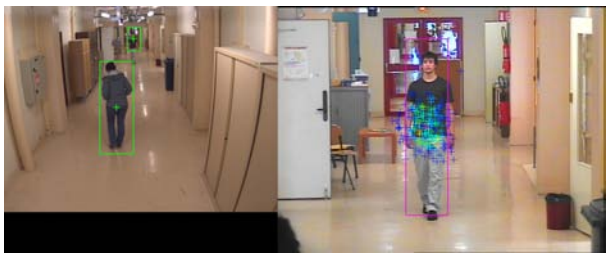


image n°111

Figure 6.6: *Suivi d'une personne avec notre système de collaboration entre la caméra statique et la caméra dynamique, cas où il y a plusieurs personnes dans la scène sans occultation.*

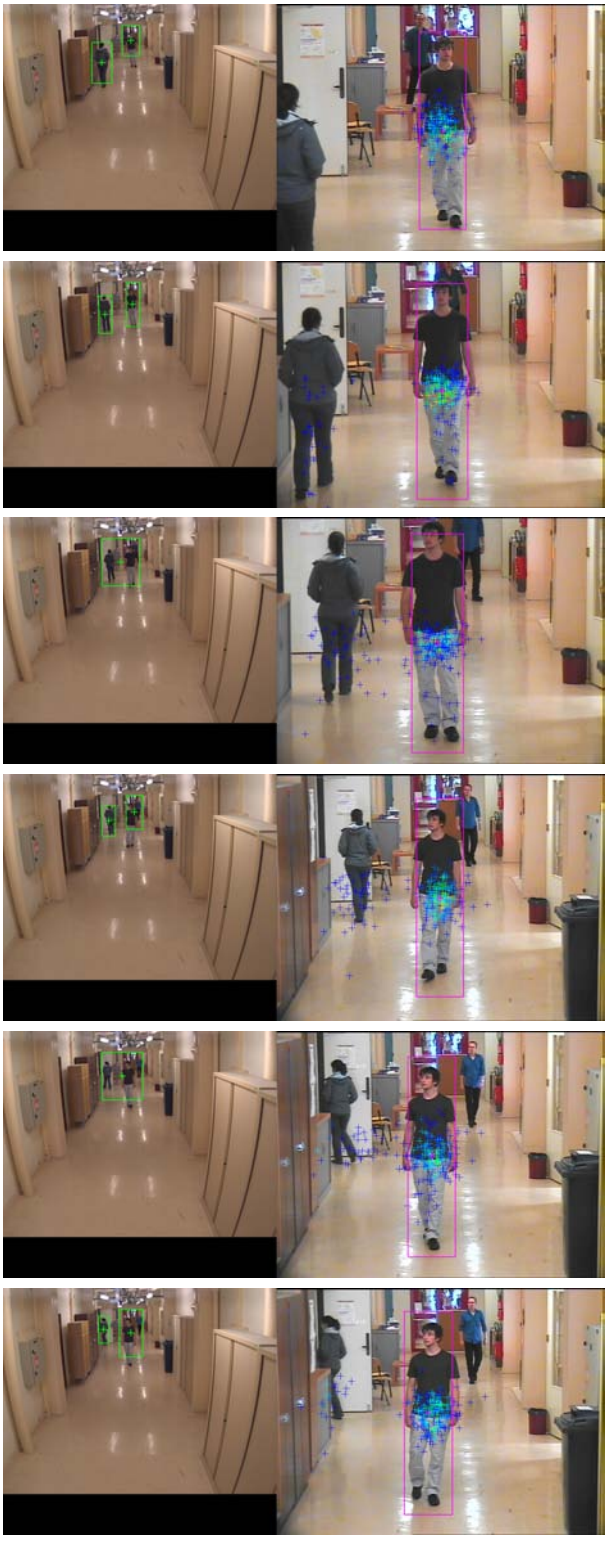


Figure 6.7: *Suivi d'une personne avec notre système de collaboration entre la caméra statique et la caméra dynamique, cas où il y a plusieurs personnes dans la scène sans occultation (suite de la figure 6.6).*

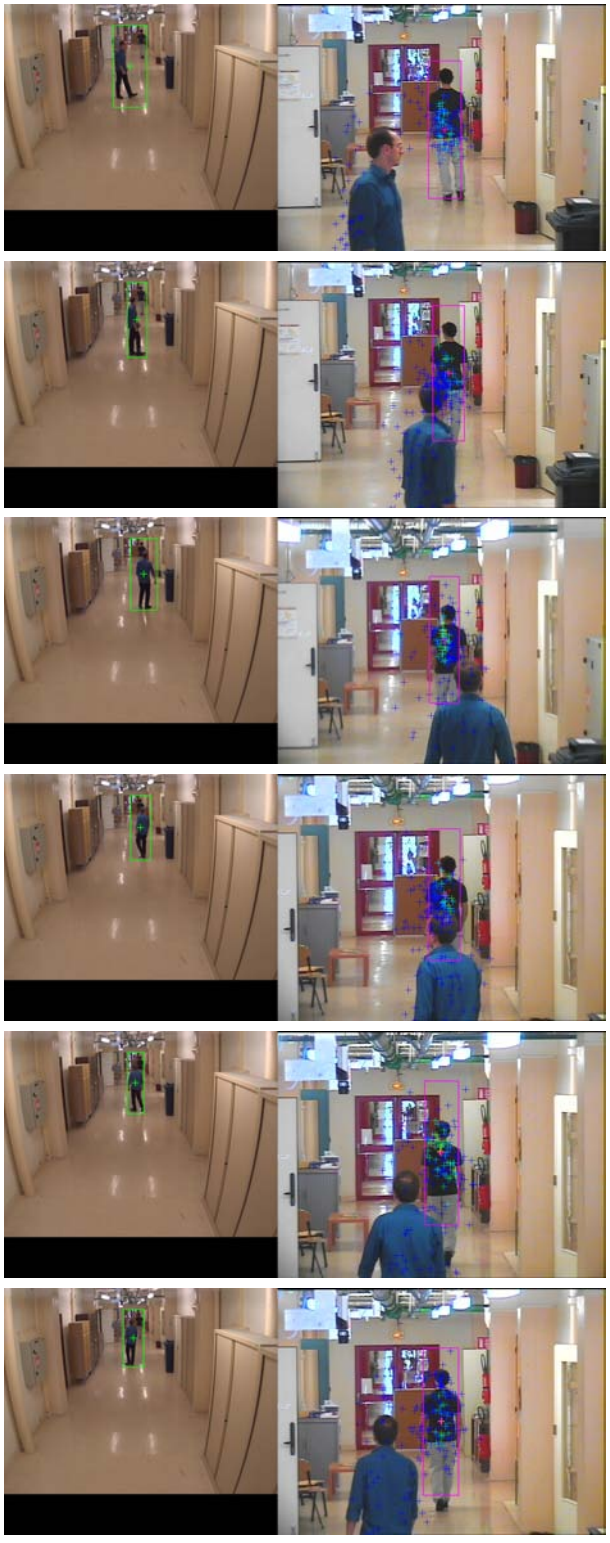


Figure 6.8: *Suivi d'une personne avec notre système de collaboration entre la caméra statique et la caméra dynamique : cas où il y a deux personnes dans la scène avec occultation partielle.*

réellement aborder des scénarii de suivi plus réalistes. Une phase d'optimisation et de parallélisation du code permettrait de se rapprocher de cet objectif.

◇ *Choix algorithmique* : Maintenant que l'on a montré au travers de ces premiers résultats l'apport d'une collaboration entre les deux caméras, on peut envisager un travail plus approfondi sur le choix des algorithmes afin d'utiliser au maximum les possibilités de la fusion de données. Une première piste de travail concerne le choix du type de descripteur d'apparence. En effet, on a choisi une modélisation basée sur l'information couleur sous la forme d'un spatiogramme (chapitre 5) qui donne de bons résultats mais qui reste sensible aux changements d'illumination. On peut envisager d'améliorer les résultats obtenus en se basant sur la primitive en utilisant les matrices de covariance [103] qui sont de plus faible dimension que les histogrammes, plus robuste et moins sensible aux changements d'illumination.
Une autre limitation de notre approche, due aux choix de la couleur comme descripteur, est la grande sensibilité de notre algorithme aux similarités d'apparence des diverses personnes présentes dans la scène : cette modélisation ne permet pas de différencier des personnes habillées de la même façon se croisant dans la scène (figure 6.9). Une piste d'amélioration est d'envisager de combiner, au descripteur basé sur la couleur, un autre descripteur basé sur le mouvement par une modélisation dynamique de la cible plus fiable ([8], [18], [54]).

◇ *Initialisation du suivi* : Actuellement, le choix de la cible et la définition du modèle sont manuelles afin que la taille du modèle soit ajustée à la personne et rendre plus performant le suivi de personne. Afin d'automatiser ce processus, on peut rendre plus robuste l'étape de détection de personne dans la caméra statique par l'utilisation d'un module de classification dédié aux objets à suivre (dans notre cas, à la détection de personne), ce qui peut permettre de raffiner la boîte englobant la cible dans l'image de la caméra statique. A partir de là, on peut envisager une interaction entre les deux caméras : on sélectionne dans l'image de la caméra statique la personne que l'on veut suivre. On estime les paramètres angulaires et de zoom à appliquer à la caméra dynamique pour se centrer sur la cible en fonction de la position et de la taille du blob. Une fois cette étape effectuée, on apprend le modèle contenue dans la zone couverte par le blob reportée dans le repère de l'image de la caméra dynamique.

Figure 6.9: *Résultat du suivi de personne basée sur la collaboration entre capteurs : la personne, suivie à la première image, recule au cours de la séquence d'images mais le processus ne suit pas son mouvement.*

Conclusion et perspectives

De nombreux travaux en vidéo-surveillance de reconnaissance et de suivi s'appuient sur des systèmes de vision monoculaire ou stéréo mais peu de recherche a été effectuée dans la mise en œuvre de systèmes hybrides. Pourtant la complexité des scènes à analyser nous invite à explorer cette voie : être capables de réagir avec une vision d'ensemble, mais également obtenir une résolution suffisante pour analyser, reconnaître, classifier les comportement des personnes dans un contexte clairement affiché de vidéo-surveillance.

Les challenges scientifiques sont nombreux : l'augmentation de la complexité du système d'acquisition s'accompagne inéluctablement de procédures de calibrage sophistiquées qui se doivent d'être précises, automatiques ou presque, robustes et rapides. De façon analogue, l'analyse d'images doit s'appuyer sur des algorithmes performants, être capable de fusionner de l'information extraite de différents capteurs à différentes résolutions et non nécessairement localisés au même point de vue.

Durant cette thèse nous avons abordé les deux volets cités précédemment et tenté d'apporter des réponses aux diverses problématiques soulevées.

Relatif au calibrage, nous avons proposé une nouvelle solution de modélisation optique et géométrique du système de vision hybride qui tend vers un dispositif automatique et autonome.

L'approche développée permet d'apprendre la relation liant dans un premier temps les coordonnées de l'image de la caméra statique et les paramètres angulaires de la caméra dynamique tels que celle-ci soit centrée sur la position définie dans la caméra statique. Cette relation est intrinsèquement sous tendue à la géométrie de la scène observée par les capteurs et donc dédiée à une scène précise. Sous une hypothèse de géométrie locale simplifiée, nous établissons une relation supplémentaire permettant d'obtenir la transformation complète liant les coordonnées d'un point dans l'image de la caméra statique aux coordonnées de ce même point dans l'image de la caméra dynamique. Les performances de la méthode de calibrage sont analysées au chapitre 3 où nous montrons le bien fondé des hypothèses dans le cadre de notre application.

Plusieurs pistes d'amélioration existent : tout d'abord, afin d'avoir une méthode de calibrage complètement automatique et autonome, il serait bon de rendre générique et automatique la phase d'étalonnage spécifique mise en place actuellement pour obtenir la relation liant les paramètres angulaires aux coordonnées dans l'image de la caméra dynamique (chapitre 2). Ensuite, nous pouvons relâcher les hypothèses de géométrie locale

planaire et de ce fait remplacer l'appariement à base d'homographie par une surface d'ordre supérieur. Enfin, si l'on souhaite un système de calibrage automatique complet, il semble intéressant de se pencher sur un module évaluant si la configuration de scène a changé ou si les caméras ont bougé et permettant ainsi de déterminer à quel moment il est nécessaire de relancer le calibrage du système.

Dans un second temps, après avoir validé notre solution de calibrage par la mise en place d'un système maître-esclave, nous avons proposé un système de suivi basé sur une collaboration étroite entre les deux caméras qui améliorent significativement les résultats de suivi. La collaboration intervient à l'étape de prédiction des particules selon la fonction d'importance q dans l'algorithme de suivi basé sur le filtre à particules utilisé dans la caméra statique. Classiquement, la fonction d'importance q est fonction du résultat du résultat de l'itération précédente du filtre. Afin d'affiner la propagation des particules en fonction des zones en mouvement détectés par la caméra statique, nous avons défini q fonction du résultat du filtre à particules sur la caméra dynamique de l'instant précédent et des détections d'objet en mouvement issues de la caméra statique (équation 5.16 du chapitre 2). Au cours du chapitre 6, nous avons montré l'apport de la collaboration entre capteurs par rapport à un schéma maître-esclave pour effectuer un suivi de personne haute résolution.

Ces premiers résultats encouragent à poursuivre le développement d'une solution de suivi basé sur cette configuration de caméras. Comme cela a été évoqué dans le dernier chapitre, les améliorations possibles du système concernent une optimisation des temps de calcul et la recherche de techniques permettant de mieux discriminer la cible. Mais le problème scientifique crucial sera de clairement définir la meilleure façon d'obtenir une information plus riche issue de la fusion d'informations hétérogènes, au sein d'un dispositif temps réel. Nous sommes en phase actuellement d'expérimentation de notre premier prototype qui devra s'enrichir de l'expérience acquise à travers sa mise en œuvre sur des scènes et des situations variées.

Publications

Chapitre de livre

J. Badri, C. Tilmant, J.-M. Lavest, Q.-C. Pham et P. Sayd. - Pattern Recognition : Techniques, Technology and Applications. - *Chapitre : Automatic calibration of hybrid dynamic vision system for high resolution object tracking* . 2008, I-Tech Education and Publishing, Vienna, Austria.

Congrès internationaux

J. Badri, C. Tilmant, J.-M. Lavest, Q.-C. Pham et P. Sayd. - Camera-to-Camera Mapping for Hybrid Pan-Tilt-Zoom Sensors calibration. - *Scandinavian Conference on Image Analysis*, SCIA'2007, Aalborg, Danemark.

J. Badri, C. Tilmant, J.-M. Lavest, Q.-C. Pham et P. Sayd. - Hybrid Dynamic Sensors Calibration From Camera-to-Camera Mapping : an Automatic Approach. - *Dans Second International Conference on Computer Vision Theory and Applications*, VISAPP'2007, Barcelone, Espagne.

Worshops et symposiums internationaux

J. Badri, C. Tilmant, J.-M. Lavest, Q.-C. Pham et P. Sayd. - Hybrid sensors calibration : Application to pattern recognition and tracking. - *IEEE International Symposium on Intelligent Signal Processing*, WISP'2007, Madrid, Espagne.

Congrès et workshops nationaux

J. Badri, C. Tilmant, J.-M. Lavest, Q.-C. Pham et P. Sayd. - Mise en relation caméra à caméra pour le calibrage de capteurs hétérogènes avec une caméra Pan-Tilt-Zoom. - *Congrès francophone des jeunes chercheurs en vision par ordinateur*, ORASIS'2007, Obernai, France.

Annexe

Caractéristiques de la caméra statique AXIS 221 : [1]

Illustration de la caméra AXIS 221.

Capteur d'images	Capteur CCD RVB à balayage progressif 1/3" Sony Wfine
Objectif	varifocale F1.0 3.0 - 8.0 mmm Auto iris Mise au point : 0.3 m à l'infini
Angle de prise de vue	35° - 93° horizontal
Compression vidéo	Motion JPEG MPEG-4 Partie 2 (ISO/IEC 144496-2) Profil ASP et SP
Résolutions	16 résolutions de 640×480 pixels à 160×120 pixels

1. Les caractéristiques complètes de la caméra AXIS 221 se trouvent sur le site de la société à l'adresse http://www.axis.com/products/cam_221/index.htm.

Caractéristiques de la caméra dynamique AXIS 213PTZ : [2]

Illustration de la caméra AXIS 213PTZ.

Capteur d'images	Capteur CCD à balayage entrelacé 1/4"
Objectif	F1.6 - 4.0 Objectif motorisé f = 3,5 - 91mm Mise au point automatique
Angle de prise de vue	$1,7°$ - $47°$ horizontal
Zoom	26× (optique) et 12× (numérique)
Rotation	$\pm 170°$
Inclinaison	$-10°$ à $90°$
Vitesse	Rotation : de 1 à 90°/sec Inclinaison : de 1 à 70°/sec
Compression vidéo	Motion JPEG MPEG-4 Partie 2 (ISO/IEC 144496-2) Profil ASP et SP
Résolutions	4CIF, 2CIFExp, 2CIF, CIF, QCIF Max. 704×480 (NTSC), 704×576 (PAL) Min. 160×120 (NTSC), 176×144 (PAL)

2. Les caractéristiques complètes de la caméra AXIS 213PTZ se trouvent sur le site de la société à l'adresse http://www.axis.com/products/cam_213/index.htm.

Caractéristiques de la caméra dynamique AXIS 233D : [3]

Illustration de la caméra AXIS 233D.

Capteur d'images	Capteur CCD à balayage progressif ExView HAD 1/4"
Objectif	F1.4 - 4.2 f = 3,4 - 119mm Mise au point automatique Profondeur de champ : 100mm (largeur) ou 1000 mm (télé.) à l'infini
Angle de prise de vue	1,73° - 55,8° horizontal
Zoom	35× (optique) et 12× (numérique). Total : 420× (optique)
Rotation	360° intégral, sans butée
Inclinaison	180°
Vitesse	Rotation : de 0,05 à 450°/sec Inclinaison : de 0,05 à 450°/sec
Compression vidéo	Motion JPEG MPEG-4 Partie 2 (ISO/IEC 144496-2) ASP niveau 0-5, SP niveau 0-3
Résolutions	4CIF, 2CIFExp, 2CIF, CIF, QCIF Max. 704×480 (NTSC), 704×576 (PAL) Min. 176×120 (NTSC), 176×144 (PAL)

3. Les caractéristiques complètes de la caméra AXIS 233D se trouvent sur le site de la société à l'adresse http://www.axis.com/products/cam_221/index.htm.

Bibliographie

[1] S. Arulampalam, S. Maskell, N. Gordon et T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking . *IEEE Transactions on Signal Processing*, 50 :174–188, 2002.

[2] Y. Bar-Shalom et E. Fortmann. *Tracking and data association*. Academic Press Professional, Inc., 1988.

[3] J. Barreto, P. Peixoto, J. Batista et H. Araujo. Tracking multiple objects in 3D . *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1 :210–215, 1999.

[4] J. L. Barron, D. J. Fleet et S. S. Beauchemin. Performance of optical flow techniques . *International journal of computer vision*, 12 :43–77, 1994.

[5] A. Basu et K. Ravi. Active camera calibration using pan, tilt and roll . *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 27 :559–566, 1997.

[6] S. S. Beauchemin et J. L. Barron. The computation of optical flow . *ACM Computing Surveys (CSUR)*, 27 :433–466, 1995.

[7] P. J. Besl et N. D. McKay. A Method for Registration of 3-D Shapes . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14 :239–256, 1992.

[8] E. Bichot, L. Mascarilla et P. Courtellemont. A new importance sampling scheme based on motion segmentation in particle filtering . *International Conference on Signal Processing, Computational Geometry & Artificial Vision*, 2005.

[9] S. T. Birchfield et S. Rangarajan. Spatiograms versus Histograms for Region-Based Tracking . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2 :1158–1163, 2005.

[10] S. T. Birchfield et S. Rangarajan. Spatial Histograms for Region-Based Tracking . *ETRI Journal*, 25 :697–699, 2007.

[11] J. Black et T. Ellis. Multi camera image tracking . *Image and Vision Computing*, 24 :1256–1267, 2006.

[12] T. Blaszka. *Approches par modèles en vision précoce* . Thèse de doctorat, Université de Nice Sophia-Antipolis, 1997.

[13] R. Bodor, R. Morlok et N. Papanikolopoulos. Dual-camera system for multi-level activity recognition . *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1 :643–648, 2004.

[14] F. L. Bookstein. Principal Warps : Thin-Plate Splines and the Decomposition of Deformations . *IEEE Transactions Pattern Analysis Machine Intelligence*, 11 :567–585, 1989.

[15] K. Bowyer, C. Kranenburg et S. Dougherty. Edge detector evaluation using empirical ROC curves . *Computer Vision and Image Understanding*, 84 :77–103, 2001.

[16] P. Brand. *reconstruction tridimensionnelle à partir d'une caméra en mouvement : de l'influence de la précision* . Thèse de doctorat, Université Claude Bernard à Lyon I, 1995.

[17] T. Broida et R. Chellappa. Estimation of object motion parameters from noisy images . *IEEE Transactions Pattern Analysis Machine Intelligence*, 8 :90–99, 1986.

[18] D. J. Bullock et J. S. Zelek. Real-time tracking for visual interface applications in cluttered and occluding situations . *Image and Vision Computing*, 22 :1083–1091, 2004.

[19] Q. Cai et J. K. Aggarwal. Tracking Human Motion in Structured Environments Using a Distributed-Camera System . *IEEE Transactions Pattern Analysis Machine Intelligence*, 21 :1241–1247, 1999.

[20] F. Chang, C.-J. Chen et C.-J. Lu. A linear-time component-labeling algorithm using contour tracing technique . *Computer Vision and Image Understanding*, 93 :206–220, 2004.

[21] T.-H. Chang et S. Gong. Tracking Multiple People with a Multi-Camera System . *Workshop on Multi-Object Tracking*, 2001.

[22] Y. Chen et G. Medioni. Object modelling by registration of multiple range images . *IEEE International Conference on Robotics and Automation*, 1991.

[23] Y.-T. Chen, C.-S. Chen, C.-R. Huang et Y.-P. Hung. Efficient hierarchical method for background subtraction . *Pattern Recognition*, 40 :2706–2715, 2007.

[24] R. Collins et Y. Tsin. Calibration of an outdoor active camera system . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1 :528–534, 1999.

[25] D. Comaniciu et P. Meer. Mean Shift : A Robust Approach Toward Feature Space Analysis . *IEEE Transactions Pattern Analysis Machine Intelligence*, 24 :603–619, 2002.

[26] D. Comaniciu, V. Ramesh et P. Meer. Kernel-Based Object Tracking . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 :564–575, 2003.

[27] C. Ó'Conaire, N. E. O'Connor et A. F. Smeaton. An improved spatiogram similarity measure for robust object localisation . *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[28] I. J. Cox. A review of statistical data association techniques for motion correspondence . *International Journal of Computer Vision*, 10 :53–66, 1993.

[29] J. Davis et X. Chen. Calibrating pan-tilt cameras in wide-area surveillance networks . *IEEE International Conference on Computer Vision*, 1 :144–149, 2003.

[30] D. Dubois et H. Prade. Possibility theory and data fusion in poorly informed environments . *Control Engineering Practice*, 1994.

[31] A. M. Elgammal, D. Harwood et L. S. Davis. Non-parametric Model for Background Subtraction . *Proceedings of the 6th European Conference on Computer Vision-Part II*, 2000.

[32] M. A. Fischler et R. C. Bolles. Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography . *Communications of the ACM*, 24 :381–395, 1981.

[33] F. Fleuret, J. Berclaz, R. Lengagne et P. Fua. Multi-Camera People Tracking with a Probabilistic Occupancy Map . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30 :267–282, 2007.

[34] S. N. Fry, M. Bichsel, P. Muller et D. Robert. Tracking of flying insects using pan-tilt cameras . *Journal of Neuroscience Methods*, 101 :59–67, 2000.

[35] A. Gardel. *Calibration of a Zoom Lens Camera with Pan&Tilt Movement for Robotics* . Thèse de doctorat, Université Blaise Pascal, Clermont-Ferrand, 2004.

[36] A. Gilbert et R. Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity . *European Conference on Computer Vision*, 2006.

[37] N. Gordon, D. Salmond et A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation . *Radar and Signal Processing*, 140 :107–113, 1993.

[38] L. Grewe et A. C. Kak. Interactive learning of a multiple-attribute hash table classifier for fast object recognition . *Computer Vision and Image Understanding*, 61 :387–416, 1995.

[39] B. Han, S.-W. Joo et L. Davis. Probabilistic Fusion Tracking Using Mixture Kernel-Based Bayesian Filtering . *IEEE International Conference on Computer Vision*, 2007.

[40] R. M. Haralick, K. Shanmugam et Dinstein. Textural Features for Image Classification . *IEEE Transactions on Systems, Man and Cybernetics*, 3 :610–621, 1973.

[41] C. Harris et M. Stephens. A combined corner and edge detector . *Alvey Vision Conference*, 1988.

[42] R. Hartley et A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.

[43] R. Horaud, D. Knossow et M. Michaelis. Camera cooperation for achieving visual attention . *Machine Vision Application*, 16(6) :1–2, 2006.

[44] R. Horaud et O. monga. *Vision par ordinateur : outils fondamentaux* , Chapitre Géométrie et calibration des caméras, page 139–186. Hermes Science Publications, 1995.

[45] W. Hu, T. Tan, L. Wang et S. Maybank. A survey on visual surveillance of object motion and behaviors . *IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews*, 34 :334–352, 2004.

[46] C.-R. Huang, C.-S. Chen et P.-C. Chung. Contrast Context Histogram - A Discriminating Local Descriptor for Image Matching . *Proceedings of the 18th International Conference on Pattern Recognition*, 2006.

[47] C. Hue. *Méthodes séquentielles de Monte-Carlo pour le filtrage non linéaire multi-objets dans un environnement bruité. Applications au pistage multi-cibles et à la trajectographie d'entités dans des séquences d'images 2D* . Thèse de doctorat, Université de Rennes I, 2003.

[48] M. Isard et A. Blake. Contour Tracking by Stochastic Propagation of Conditional Density . *European Conference on Computer Vision*, 1996.

[49] M. Isard et A. Blake. ICONDENSATION : Unifying Low-Level and High-Level Tracking in a Stochastic Framework . *European Conference on Computer Vision*, 1998.

[50] A. Jain, D. Kopell, K. Kakligian et Y.-F. Wang. Using Stationary-Dynamic Camera Assemblies for Wide-area Video Surveillance and Selective Attention . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1 :537–544, 2006.

[51] R. Jain et H. Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes . *IEEE Transactions Pattern Analysis and Machine Intelligence*, 1 :206–214, 1979.

[52] T. Jost. *Fast geometric matching for shape registration* . Thèse de doctorat, Faculté des sciences de l'université de Neuchâtel, 2002.

[53] T. Kailath, A. Sayed et B. Hassibi. *Linear estimation.* Prentice Hall, 2000.

[54] J. Kang, I. Cohen et G. Medioni. Continuous tracking within and across camera streams . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1 :267–272, 2003.

[55] S. Khan et M. Shah. Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 :1355– 1360, 2003.

[56] K. Kim et L. S. Davis. Multi-camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering . *European Conference on Computer Vision*, 2006.

[57] G. Kitagawa. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models . *Journal of Computational and Graphical Statistics*, 5 :1–5, 1996.

[58] J.-M. Lavest et G. rives. *Perception visuelle par imagerie vidéo* , Chapitre Etalonnage des capteurs de vision, page 23–58. Hermes Science Publications, 2003.

[59] J. Lavest et M. Dhome. Quelle précision pour une mire d'étalonnage ? . *Traitement du Signal*, 16 :241–246, 1999.

[60] K. Levenberg. A Method for the Solution of Certain Problems in Least Squares . *Quart. Appl. Math*, 2 :164–168, 1944.

[61] A. J. Lipton, H. Fujiyoshi et R. S. Patil. Moving Target Classification and Tracking from Real-time Video . *IEEE Workshop on Applications of Computer Vision*, page 8, 1998.

[62] R. Longchamp. *Commande numérique de systèmes dynamiques : Cours d'automatique*. Presses Polytechniques et Universitaires Romandes, 2006.

[63] D. G. Lowe. Object Recognition from Local Scale-Invariant Features . *IEEE International Conference on Computer Vision*, 02 :1150, 1999.

[64] R. Luo, C.-C. Yih et K. L. Su. Multisensor fusion and integration : approaches, applications, and future research directions . *IEEE Sensors Journal*, 2 :107–119, 2002.

[65] C. Madden, E. D. Cheng et M. Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation . *Machine Vision and Applications*, 18 :233–247, 2007.

[66] S. Mallat. A Theory for Multiresolution Signal Decomposition : The Wavelet Representation . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11 :674–693, 1989.

[67] E. Marchand et F. Chaumette. A new formulation for non-linear camera calibration using virtual visual servoing . Rapport de recherche, IRISA, No 1366, 2001.

[68] X. Marichal, W.-Y. Ma et H. Zhang. Blur determination in the compressed domain using DCT information . *International Conference on Image Processing*, 2 :386–390, 1999.

[69] M. Mason et Z. Duric. Using histograms to detect and track objects in color video . *Applied Imagery Pattern Recognition Workshop*, 2001.

[70] K. Mikolajczyk et C. Schmid. A Performance Evaluation of Local Descriptors . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10) :1615–1630, 2005.

[71] A. Mitiche et P. Bouthemy. Computation and analysis of image motion : a synopsis of current problems and methods . *International journal of computer vision*, 19 :29–55, 1996.

[72] A. Monnet, A. Mittal, N. Paragios et V. Ramesh. Background Modeling and Subtraction of Dynamic Scenes . *IEEE International Conference on Computer Vision*, 2003.

[73] P. D. Moral, G. Rigal et G. Salut. Estimation et commande optimale non-lineaire : un cadre unifié pour la resolution particulaire . Rapport de fin de contrat dret, LAAS, Toulouse, 1992.

[74] C. Motamed et O. Wallart. A temporal fusion strategy for cross-camera data association . *Pattern Recognition Letters*, 28 :233–245, 2007.

[75] Y. Nam, J. Ryu, Y.-J. Choi et W.-D. Cho. Learning Spatio-Temporal Topology of a Multi-Camera Network by Tracking Multiple People, . *Journal of Signal processing*, 2007.

[76] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little et D. G. Lowe. A Boosted Particle Filter : Multitarget Detection and Tracking . *European Conference on Computer Vision*, 1 :28–39, 2004.

[77] N. Otsu. AA threshold selection method from grey scale histogram . *IEEE Transactions on Systems Man and Cybernetics*, 1979.

[78] C. Papageorgiou, T. Evgeniou et T. Poggio. A trainable pedestrian detection system . *IEEE Intelligent Vehicles Symposium*, 1998.

[79] E. Parzen. On the estimation of a probability density function and mode . *Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[80] B. Peuchot. Utilisation de détecteurs subpixels dans la modélisation d'une caméra. . *RFIA*, 1994.

[81] Q.-C. Pham, Y. Dhome, L. Gond et P. Sayd. Video Monitoring of Vulnerable People in Home Environment . *Conference On Smart Homes and Health Telematics*, 2008.

[82] M. Piccardi. Background subtraction techniques : a review. . *IEEE international Conefrence on Systems, Man and Cybernetics*, 2004.

[83] F. Porikli. Achieving real-time object detection and tracking under extreme conditions . *Journal of Real-Time Image Processing*, 1 :33–40, 2006.

[84] F. Porikli et O. Tuzel. Fast Construction of Covariance Matrices for Arbitrary Size Image Windows . *IEEE International Conference on Image Processing*, page 1581–1584, 2006.

[85] F. Porikli et O. Tuzel. Bayesian background modeling for foreground detection . *ACM international workshop on Video surveillance*, 2005.

[86] F. Porikli, O. Tuzel et P. Meer. Covariance Tracking using Model Update Based on Lie Algebra . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[87] P. Pérez. Modèles et algorithmes pour l'analyse probabiliste des images . *Habilitation à diriger des recherches de l'Université de Rennes 1*, 2003.

[88] P. Pérez, J. Vermaak et A. Blake. Data Fusion for Visual Tracking With Particles . *Proceedings of the IEEE*, 2004.

[89] J. Rittscher, J. Kato, S. Joga et A. Blake. A Probabilistic Background Model for Tracking . *Proceedings of the 6th European Conference on Computer Vision-Part II*, 2000.

[90] H. A. Rowley, S. Baluja et T. Kanade. Neural Network-Based Face Detection . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 :23–38, 1998.

[91] A. W. Senior, A. Hampapur et M. Lu. Acquiring Multi-Scale Images by Pan-Tilt-Zoom Control and Automatic Multi-Camera Calibration . *Workshops on Application of Computer Vision*, 1 :433–438, 2005.

[92] J. Shi et J. Malik. Normalized Cuts and Image Segmentation . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :888–905, 2000.

[93] J. Shi et C. Tomasi. Good features to track . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1994.

[94] H.-Y. Shum, M. Han et R. Szeliski. Interactive Construction of 3D Models from Panoramic Mosaics . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 427, 1998.

[95] R. Sprengel, K. Rohr et H. S. Stiehl. Thin-plate spline approximation for image registration . *IEEE Engineering in Medicine and Biology Society*, 3 :1190–1191, 1996.

[96] C. Stauffer et W. Grimson. Adaptive Background Mixture Models for Real-Time Tracking . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2 :2246, 1999.

[97] C. Stauffer et K. Tieu. Automated multi-camera planar tracking correspondence modeling . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1 :259, 2003.

[98] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee et J. M. Buhmann. Topology Free Hidden Markov Models : Application to Background Modeling . *IEEE International Conference on Computer Vision*, 1 :294, 2001.

[99] H. Tanizaki. Non-gaussian state-space model of nonstationary time series . *Journal of the American Statistical Association*, 1987.

[100] K. Tieu, G. Dalley et W. E. L. Grimson. Inference of Non-Overlapping Camera Network Topology by Measuring Statistical Dependence . *IEEE International Conference on Computer Vision*, 2005.

[101] H. Tong, M. Li, H.-J. Zhang et C. Zhang. Blur Detection for Digital Images Using Wavelet Transform . *International Conference on Multimedia and Expo*, 2004.

[102] K. Toyama, J. Krumm, B. Brumitt et B. Meyers. Wallflower : Principles and practice of background maintenance . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 255–261, 1999.

[103] O. Tuzel, F. Porikli et P. Meer. Region Covariance : A Fast Descriptor for Detection and Classification . *European Conference on Computer Vision*, 2006.

[104] C. J. Veenman, M. J. T. Reinders et E. Backer. Resolving Motion Correspondence for Densely Moving Points . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 :54–72, 2001.

[105] P. Viola et M. Jones. Rapid object detection using a boosted cascade of simple features . *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1 :511–518, 2001.

[106] P. Viola, M. J. Jones et D. Snow. Detecting Pedestrians Using Patterns of Motion and Appearance . *IEEE International Conference on Computer Vision*, 2003.

[107] D. Woo et D. Capson. 3D visual tracking using a network of low-cost pan/tilt cameras . *Canadian Conference on Electrical and Computer Engineering*, 2 :884–889, 2000.

[108] C. R. Wren, A. Azarbayejani, T. Darrell et A. Pentland. Pfinder : Real-Time Tracking of the Human Body . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 :780–785, 1997.

[109] Y. Yao, B. Abidi et M. Abidi. Fusion of Omnidirectional and PTZ Cameras for Accurate Cooperative Tracking . *IEEE International Conference on Video and Signal Based Surveillance*, 2006.

[110] A. Yilmaz, X. Li et M. Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 :1531–1536, 2004.

[111] A. Yilmaz, O. Javed et M. Shah. Object tracking : A survey . *ACM Computing Surveys (CSUR)*, 38 :13, 2006.

[112] X. Zhou, R. T. Collins, T. Kanade et P. Metes. A master-slave system to acquire biometric imagery of humans at distance . *ACM international workshop on Video surveillance*, 0 :113–120, 2003.

Résumé

La vidéo surveillance est un sujet scientifiquement complexe dès lors que l'on souhaite intégrer un degré d'automatisation élevé tant dans la détection et la reconnaissance d'un objet spécifique présent dans la scène, que dans la structure de commande rapide et précise du dispositif d'acquisition. Le cadre d'étude de cette thèse se place dans le contexte du suivi automatique de personne à travers la mise en œuvre d'un système de vision hybride (caméra statique et dynamique). Si la littérature est riche dans les dispositifs de vision orientables, les travaux relatifs à la collaboration multi-capteurs sont plus rares dans ce domaine d'application, qui plus est pour des capteurs hybrides alliant un grand champ de vue et une résolution faible avec un dispositif Pan-Tilt-Zoom (PTZ) orientable garantissant le détail de l'information.

Le manuscrit aborde plusieurs problématiques induites par le sujet et donne de premiers résultats de suivi. Une part importante est consacrée à la caractérisation optique et géométrique du dispositif de vision et à la proposition d'une solution originale de calibrage permettant d'estimer une fonction de transfert lorsque les centres optiques des deux capteurs sont pratiquement confondus. L'approche décrite permet d'inférer automatiquement la scène tridimensionnelle observée par le dispositif. Elle établit une relation d'asservissement entre une zone d'intérêt détectée dans l'image grand champ et la commande à appliquer sur le dispositif PTZ pour se focaliser.

Dans une seconde partie, nous abordons la mise en relation des informations extraites de chaque capteur et leur enrichissement mutuel nécessaire à la réalisation du suivi de personne. Après une étude bibliographique, nous détaillons la mise en œuvre d'une approche de filtrage particulaire. L'étape de prédiction du filtre est guidée par la détection issue de la caméra statique et la mesure est donnée par une modélisation d'apparence de la cible extraite de la caméra dynamique. Le formalisme complet du filtre à particules inclut également la loi de commande de la caméra PTZ. Enfin, les premiers résultats de suivi du système complet sont présentés et analysés.